Collective Sanctions:

An experimental approach

Philipp Chapkovskii

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Political and Social Sciences
of the European University Institute

14 September 2018

Florence

European University Institute
Department of Political and Social Sciences

Collective Sanctions:

An experimental approach

Philipp Chapkovskii

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Political and Social Sciences
of the European University Institute

Examining Board

Supervisor Prof. Diego Gambetta, European University Institute (EUI)
Prof. Elias Dinas, European University Institute (EUI)
Prof. Andreas Flache, University of Groningen
Prof. Siegwart Lindenberg, Tilburg University

Researcher declaration to accompany the submission of written work
Department of Political and Social Sciences - Doctoral Programme

I Philipp Chapkovskii certify that I am the author of the work "Collective Sanctions: an experimental approach" I have presented for examination for the Ph.D. at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

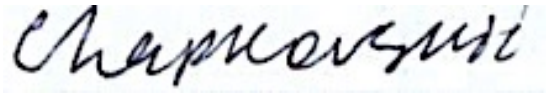
I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the Code of Ethics in Academic Research issued by the European University Institute (IUE 332/2/10 (CA 297).

I declare that this work consists of 33.233 words.

Statement of language correction:
This thesis has been corrected for linguistic and stylistic errors. I certify that I have checked and approved all language corrections, and that these have not affected the content of this work.

Signature and date:



Abstract

Collective Sanctions: an experimental approach

by Philipp Chapkovskii

This dissertation tests the efficiency of collective sanctions as a preventive measure experimentally with three different cases: (i) collective sanctions and the production of public good; (ii) collective sanctions and intergroup cooperation, and (iii) collective sanctions and peer punishment. The dissertation demonstrates that in all three cases the functionalist hypothesis of a potential efficiency of collective sanctions does not find empirical support.

In the first chapter, I test if sanctions applied to an entire group for the free-riding of one of its members can increase the level of cooperation within that group. To measure the efficiency of such collective sanctions, I conducted a lab experiment based on a standard public good game. The results show that overall, collective sanctions are inefficient. Moreover, when subjects are able to punish their peers, the level of cooperation is lower in the regime of collective sanctions than under individual sanctions.

The second chapter tests whether collective sanctions applied by outgroup members result in higher intergroup cooperation, and whether the introduction of collective sanctions increases the amount of ingroup punishment. The results demonstrate that neither of these two functionalist arguments come true: participants avoid using collective sanctions against outgroups, and the amount of ingroup (third-party) punishment is no higher under the intergroup collective sanctions regime. As a result, the introduction of intergroup collective sanctions does not result in the higher degree of intergroup cooperation.

The third chapter analyzes how the introduction of collective sanctions affects the willingness to punish norm violations. I conducted a lab experiment in which participants can choose to take money from a charity. After having taken their own decision, they can observe the decisions of others, and can decide to punish them. The results demonstrated that collective sanctions significantly increased the frequency of peer punishment. However, this increased rate of punishment did not go along with lower crime rates.

Contents

Introduction

In 2002, the Russian margarine and oil producing company Efko, which owns vast territories of agricultural land in southern Russia, introduced a new motivational system for farmers. Efko decided to institute this system to motivate rural Russian farmers to work more efficiently and with better morale. The system's main characteristic was a focus on collective liability of workers; it was introduced after running a series of sociological and psychological tests in Russian villages. Workers' efficiency was measured at the group level; any kind of work ethics violation (e.g. absenteeism, working while drunk, destruction of tools, etc.) brought punishment on the entire team, not just the individual. Thus, in concrete terms, all rule violations reduced the earnings of the entire group of farmers.

In an interview for a business journal, the CEO of Efko explained the logic behind these practices:

"*We try to create socio-economic relations that would include a person into the collective. Of course, a peasant should get enough money for normal life. But at the same time, people around him should be dependent on the result of his labor. The most effective guarantee of his efficiency is not the amount of money he gets, but the reaction of others. As soon as I start working badly, that affects everybody around me. And that's the factor that guarantees my efficiency much more than money. For my neighbor Vassia, what matters is to know that I care about Vassia's wellbeing. And I know that if I don't care about his wellbeing he will take an awl and will correct me. Our system is a mix of individualism and interdependency, of checks and balances.*" (Hisamova, 2002).

This quote is highly illustrative of the main points that I am going to make in the course of this thesis. I focus on the mechanism that moves the emphasis from the individual consequences of someone's actions towards a group-based incentivization tool. What happens if one's individual actions may have consequences on the well-being of others?

According to current literature, we define the transfer of negative consequences for individual actions to the entire group as collective sanctions (Heckathorn, 1988; Levinson, 2003), collective liability (Pettit, 2007), or collective punishment (Pereira et al., 2015). These terms are largely interchangeable. The definition of collective sanctions is seen as intuitively clear, so much so that one of the two existing sociological papers on this topic does not even provide a formal definition of the phenomenon (Heckathorn, 1988). The second paper uses a definition borrowed from the pedagogical manual "Strategies for Managing Behavior Problems in the Classroom," in which collective sanctions are defined as a practice in which "consequences are applied to the group contingent upon each member reaching a specified level of performance" (Whitmeyer, 2002). A recent psychological paper that studies different student groups' reactions to collective punishment for plagiarism defined collective punishment as "the negative treatment inflicted by authorities or by an outgroup upon an entire social group, in reaction to an offense committed by one or some of its members" (Pereira et al., 2015). This thesis employs a definition given by the legal scholar Daryl Levinson in his comprehensive review of the legal side of collective sanctions: "Sanctions are collective when they are threatened against or imposed on groups of two or more individuals" (Levinson, 2003).

Collective sanctioning for individual norm violations is perceived by many as a hallmark of authoritarian or colonial regimes. Whittaker (2015) said, "The colonial authorities believed that the use of collective punishment was

an appropriate method of dealing with its African population due to a belief that stock theft was a socially accepted form of accumulation in African societies. Policing and punishment could therefore be extended to the family, village or entire location of the individual(s) implicated in a crime" (645). Collective sanctions were also widely used by the Nazis in their crackdown against local populations during World War II (Mannheim, 2013, p.50). The UK authorities used their Collective Punishment Ordinance against the Kenyan population (Ibid., p. 49). Collective punishment was also applied by the Israeli authorities in Palestine and by Pakistani authorities in Bangladesh (Horvitz and Catherwood, 2009, p.89-90). In the Soviet Gulag, the underperformance of one member of a team negatively affected his peers' evaluation, resulting in decreased food rations for the team as a whole. This system of collective responsibility (Russian: krugovaia poruka, literally frankpledge) "extended social control beyond where Soviet authority could reach, giving every member of a brigade incentive to keep their fellow prisoners in line" (Barnes, 2011, p.80).

Indiscriminate violence and disjoint norms

Collective sanctions imposed upon a certain group by a state or a state-like authority usually are not intended to propagate a specific norm for the benefit of a group, but rather to control and subdue a subordinate population. When the authorities prefer such a sanctioning regime to more subtle and less provocative individual sanctions? Here it is important to distinguish between disjoint and conjoint norms.

For collective sanctions to exist it is necessary to have at least two actors, and at least one of them should be a collective entity. An actor external to this entity should be the one who is authorized to apply sanctions to an entire group. Since the collective sanctions are applied by someone who is not a target of the norms himself, in many cases the norms being enforced are disjoint

norms (Coleman, 2000). The fact that a target and a source of sanctions are unmatched, is one of the potential sources of failure of collective sanctions: the norm these sanctions enforce can be non-beneficial for a group. If an external group imposes the control through violence it is not technically correct to say that we are talking about norm establishment but rather a compliance to the rule. An important case of such disjoint norm enforcement is when violence is used by military groups upon the population living in a disputed territory. In order to curb the resistance, detect and eliminate renegades and stealthy supporters of an enemy belligerents often use violence over entire villages instead of catching specific individuals. The seminal study (Kalyvas, 2006) has examined under what conditions the belligerent parties recourse to indiscriminate violence instead of punishing selectively. The definition of such a violence makes it similar to collective sanctions: "a type of violence whereby the victims are selected on the basis of their membership in some group and irrespective of their individual actions." (Kalyvas, 2004).

Kalyvas determines four potential reasons why an external authority may prefer using indiscriminate violence: truncated data, ignorance, institutional constraints and cost. First three explain its existence mostly by error or as an unintended by-product of other decisions (such a way how decision-making process is organized within the authority). Kalyvas believes however that the main reason for indiscriminate violence is cost: "Identifying, locating, and "neutralizing" enemies (and their civilian collaborators) one by one requires a complex and costly infrastructure." This cost that prohibits a proper individual investigation, arises according to Kalyvas from the information asymmetry: paramilitary groups who are not rooted in a territory do not know who exactly can be a defector: "Enemies can be hidden among the apparently supporters of a community, and contenders can only deal with such informational problems in efficient ways by exercising violence against previously selected defectors".

However collective punishment is not only characteristic of military or authoritarian operations against peaceful populations; we can easily observe this practice in numerous everyday practices. One obvious example is in sports: FIFA regularly imposes closed-door sanctions, denying all of a club's fans the privilege to watch a game at a stadium if one of them has committed racist offenses. For instance, one of the largest Russian soccer teams, CSKA, in 2014 received the biggest anti-racism punishment in UEFA's history when some of its fans shouted racial offences. As a result, no CSKA fans were admitted for two away games, and three home games were played in an entirely empty stadium (Ellingworth, 2014). In her comprehensive overview of collective sanctions in NFL, Peterson, 2012 provides an example of the Pittsburgh Steelers, who accumulated 101 penalties in 2011, all incurred by a single player.

Beyond sports, Peterson, 2012 also cites the US' "One Strike and You're Out" eviction policy, which gives housing authorities the right to evict an entire family if any member of the household or even a guest participated in certain criminal activities, such as drug peddling. This same type of logic appears in recent prohibition legislation in the Indian state of Bihar, where there can be a criminal case brought up against all adult members of a household if just one member is suspected of consuming alcohol ("Bihar's law" 2016).

Schools also commonly use collective sanctions. Teachers often give an entire class detention when only one student misbehaves, and, as we saw in one recent Manhattan school, an entire student body can bear the punishment. In 2010, a Manhattan principal banned all 2445 students from using toilets for the entire day because of the fight that happened the day before (Schuck, 2010). Even more examples of collective sanctions in schools can be seen in Whitmeyer, 2002 and Piaget, 1965.

In the United States, collective punishment is also used in juvenile detention facilities and military boot camps (Heckathorn, 1988). In a large, first-hand witness report of recent military practices in the US Army, the author notices, "You must act like everyone else. You must perform like everyone else. If you don't, you will be punished. Or worse, the group will suffer for your mistakes" (Mockenhaupt, 2007). The homogeneity of armed forces and soldiers' obedience to the rules are mostly achieved via group sanctions: "The threat of collective punishment for individual infractions is one of the most powerful motivators in military training" (Ibid.) Mockenhaupt tells the story of his own experience with collective sanctions in the military:

"*One night as we slept, just a few days into our training, two recruits left the barracks and walked toward town, looking for a convenience store. A drill sergeant driving home picked them up a short distance from the barracks. We were awakened, told what had happened, and told we would be dealt with later. We fell back asleep knowing the morning would bring pain. 'So you want to play games?' one of our drill sergeants said. 'OK, we will play games.' He ordered us to squat and hold out our arms. The two recruits stood in front of the formation, watching us and looking sheepish. 'Don't be mad at me; be mad at your friends standing up here,' the drill sergeant said.* $< \cdots >$ *"I want you to be pissed at your friends. They did this to you. They don't want to be part of the team.*" (Ibid.)

Collective sanctions are also used in business. The method of group motivation described by the Russian CEO above is not novel: a wide range of management literature recommends linking bonuses not to individual performance of an employee but to the team-level Key Performance Indicators (KPIs) (Kerrin and Oliver, 2002; Bamberger and Levi, 2009).

Collective sanctions have been studied academically mainly by legal scholars and psychologists and not so much by sociologists. Legal scholars typically examine the efficiency of collective sanctions in crime deterrence. In

their opinion, collective sanctions can be of practical use because team members can be "in an advantageous position to identify, monitor, and control responsible individuals, and can be motivated by the threat of sanctions to do so" (Levinson, 2003). Legal theorists have discussed whether this practice should be extended to other areas, such as leveraged sanctions to prevent corporate fraud in accounting (Velikonja, 2011) or fiduciary duty cases (Ibrahim, 2008).

Moreover, collective sanctions de facto exist in American jurisprudence: for example, the Pinkerton liability rule states that members of a criminal conspiracy are responsible for each other's crimes committed during the conspiracy. Levinson, who produced the most comprehensive legal analysis of collective sanctions in the United States, treats any form of vicarious responsibility as collective sanctions, including, for instance, corporate liability, which imposes the responsibility for the torts of corporation employees upon shareholders of the company (Levinson, 2003, p.362-371).

The strong association of collective sanctions with military actions and oppressive regimes repelled many sociologists from paying more attention to this phenomenon. That may explain the relative scarcity of sociological literature on this topic. There are two main theoretical sociological papers which analyze collective sanctions. A paper by Heckathorn, 1988 focused on group size and internal cohesion as the main predictors of the efficiency of collective sanctions. He argues that if a group is large and cohesive enough, the implementation of collective sanctions will bring about a riot against the central authority deemed to be unfair, instead of higher compliance. Later on Heckathorn (1990) returns to this topic in his follow-up paper which focuses on preferences of actors and how they affect the efficiency of different group-based incentives.

In this paper Heckathorn develops a theory of a 'group-mediated social control'. Regardless whether the incentives are negative (as it is in case of

collective punishment), or positive (group-based rewards), any regime that involves such a group-based incentive structure induces an additional peer pressure. Heckathorn with his colleagues tested the effectiveness of such peer-driven intervention on the adherence rate of HIV users (Broadhead et al., 2002) with promising positive results. This method of group-based incentives suggested by Heckathorn was later used in a series of other field experiments with mixed results. An experiment in Bolivia and Peru targeted to increase conservational efforts of local farmers has shown that collective rewards crowd out social norms aimed to protect nature without bringing in positive results Narloch, Pascual, and Drucker (2012), while similar experiments in Mexico and Tanzania demonstrated that collective rewards may work if there is a certain amount of group cohesion and trust to a group leadership Kerr, Vardhan, and Jindal (2012). Collective rewards as a mechanism of creating peer pressure seems to have similar properties as collective (negative) sanctions, but without a strongly negative reputation of the latter. From that point of view using positive group rewards would be a better fit for experimental design. On the other hand a potential or real loss suffered by entire group or by a random member as a result of imposed collective sanctions is not symmetrical to collective rewards if the main postulates of prospect theory are valid Kahneman and Tversky (2013).

The second seminal paper, by Whitmeyer, 2002, explored an optimal combination of individual and collective sanctions that would produce a higher level of compliance for a minimum cost. He discovered that an optimal solution is always a corner one: one should not mix individual and collective sanctions to attain the best result. To date, there have been only three experimental papers on collective sanctions, one of them unpublished, offering empirical evidence: Dickson, 2007, Fatas, Morales, and Ubeda, 2010, and Yefeng, Jiang, and Villeval, 2015. (see their description in more details in Chapter 1).

Psychologists have not shied away from investigating the individual perceptions of collective sanction regimes. Interest in these questions can be traced back to the 1930s when Jean Piaget studied collective sanctions in a school classroom environment. In a vignette experiment, 60 children evaluated three distinct types of collective sanctions: "1. The adult does not attempt to analyze individual guilt and punishes the whole group for the offence committed by one or two of its members. 2. The adult wants to discover the transgressor, but the latter does not own up and the group refuses to denounce him. 3. The adult wants to discover the transgressor but the latter does not own up and the group is ignorant of his identity" (Piaget, 1965, p.232). In each situation, every child had to estimate how fair it was to implement collective sanctions on the entire group and express why. Piaget found that most children did not support the idea of collective sanctions. According to Piaget, in order to support collective sanctions a person's preferences should meet two necessary conditions: the strong belief into expiatory justice ("no misdeed could remain unpunished") and group solidarity ("we are ready to be punished to bear the suffering that our mate would bear"). The younger kids (up to the age of 7 or 8) supported the first idea, but being egocentric, they lacked a feeling of group solidarity, whereas older kids (over age 8) had a strong feeling of solidarity but did not believe anymore in expiatory justice.

Piaget's experiments, as well as later studies, focused on group perceptions of collective sanctions, i.e. the question of the legitimacy and fairness of such measures. Although this approach does not measure behavioral changes due to collective sanctions, it measures attitudes towards them. The fairness of a specific sanction regime is of a crucial importance because the feeling of an unjust punishment decreases willingness to cooperate and punish others, even if this strategy would be individually profitable from the rational point of view.

The early experiments of Piaget provided a basis for more recent psychological experiments in the same vein. Pereira et al., 2015 presented to subjects several vignettes, in which a teacher punishes the whole group that worked on a project together for the plagiarism of one of the students. The experimenter asked the subjects to evaluate the fairness of such a system. In general, participants perceived it as unfair. Cushman, Durwin, and Lively, 2012 found that the legitimacy of collective sanctions is context-specific. In their study of baseball fans, they discovered that, in certain situations, (e.g. when baseball players of one team harmed a randomly chosen member of another team for something that another member of opponent's team did in the previous round) collective sanctions were considered fair.

A particular type of collective sanctions, one which interests me specifically and sparked the inquiry presented in this thesis, is the cases in which, rather than the whole group receiving punishment as a result of an individual's misdeed, a randomly chosen member of the group is punished instead. Here, it makes sense to return to Levinson's definition of collective sanctions. He mentions that: "Collective should be understood in an ex ante sense, to take account of the fact that sanctions directed against a single group member chosen at random will have the same expected disutility for all group members as sanctions divided evenly among all group" (Levinson, 2003). Thus, according to Levinson, if an individual punishes a random member of another group instead of punishing the person who is guilty in a certain harmful action, we also deal with a subtype of collective sanctions.

Social psychologists have a long history of studying this displaced aggression or vicarious retribution, a punishing behavior that targets not the perpetrator but rather someone else from the same group. We may talk about vicarious retribution when it is "directed at outgroup members who, themselves, were not the direct causal agents in the original attack against the person's ingroup" (Lickel et al., 2006). The term displaced aggression was first

developed by Dollard et al., 1939 in their study of aggression, and since then has been studied in dozens of psychological papers (see Marcus-Newhall et al. 2000 for a meta-review). As psychological studies, they focus on the individual motives and effects of such revenge, rather than on reasons and consequences such behavior may have at the group level (one of the rare exceptions from such ontological individualism is Lickel et al. 2006). Psychologists have outlined several major factors guiding the displaced revenge, including strong ingroup identification and outgroup entitativity (Stenstrom et al., 2008; Gaertner and Schopler, 1998).

This dissertation studies collective sanctions from a sociological perspective. This perspective takes the collective sanctions as granted, leaving out the question whether their application is morally or legally justified which is the subject of interest mainly for legal scholars (Levinson, 2003) or moral philosophers (French, 1987). At the same time it goes beyond purely individualistic point of view (an approach used by social psychologists such as Pereira et al. (2015) or Cushman, Durwin, and Lively (2012)), trying to observe whether the threat of collective sanctions affects cooperation, the main building block of a human group (Homans, 2017).

The guiding line of inquiry is into what kinds of microsociological mechanisms are in action when we implement collective sanctioning. I examine the effect of collective sanctioning on three different aspects of interpersonal relations: cooperation within a group, cooperation between groups, and willingness to punish peers for third-party norm violations.

Collective sanctions are a puzzling sociological phenomenon. They contradict the traditional logic of punishment. Sanctioning norm violations in general and crimes specifically is a second-order public good (Heckathorn, 1989; Fehr and Gächter, 2002a). This action results in higher cooperation and less destructive egoistic activity of each member of a community, thus it is considered beneficial for the entire society. Norm enforcement is not

free: punishment of a wrongdoer is risky, associated with potential conflict with the target of sanctioning, involves significant efforts and can be time consuming. All these considerations prevent many people from being actively involved in norm enforcement even if they benefit from the environment where norms are strictly followed. Rational choice theory predicts that no rational agent will enforce the norm, preferring that the others do that for him. However, numerous studies have shown that this is not the case (Nikiforakis, 2008; Balliet, Mulder, and Van Lange, 2011; Fehr and Gächter, 1999). People not only punish those with whom they had a direct experience of non-cooperation or norm violation, but also they enforce the norm even if they are bystanders who simply observe a norm violation that does not affect them directly (Fehr and Fischbacher, 2004; McAuliffe, Jordan, and Warneken, 2015; Lergetporer et al., 2014; Nelissen and Zeelenberg, 2009). The reasons for such an altruistic punishment can be inequality aversion, spite, or social norms (Casari, 2005). But whatever guides the punisher, his aim is to harm the target of punishment in order to change the status quo (such as the distribution of income), or to influence the future course of actions of a perpetrator. This is not the case however for collective or random sanctioning. By definition this kind of sanctioning does not affect the norm violator. Theoretically, we should observe collective sanctions extremely rarely in day-to-day life, but there is a wealth of practical and academic evidence that shows collection sanctions are widespread. In this dissertation, I examine the reasons for this paradox.

In the three papers that constitute this thesis, I analyze what kind of consequences collective sanctions may have on such group-level factors as cooperation and norm compliance. I decided to focus more on behavioral dimensions of such a sanctioning regime. The reason for that is that other important aspects such as perceived fairness or an outgroup entitativity (that is, the tendency to treat outgroup members as a non-distinguishable whole) are

covered by previous psychological studies.

I begin with the level of cooperation within groups. In Chapter 1, I use a voluntary contribution design also known as public good game (Ledyard, 1994) and adapt it for comparing the relative ingroup cooperation rate under two regimes. In a standard public good game, each individual is provided with an initial amount of money which he or she can privately consume or invest into a group project. While the inputs to this collective project depend on individual decisions, its profits are equally distributed among the group members. This creates a tension between willingness to cooperate and temptation to free ride while enjoing the benefits of inputs of others. This simple design makes it an ideal baseline scenario to study group dynamics which resulted in hundreds of studies in sociology and behavioral economics (Zelmer, 2003). Collective sanctions by definition are applied upon the group from outside: either by an external authority such as the state or by members of another group. Chapter 2 focuses on collective sanctions applied by an external group (in contrast to the sanctions applied 'from above' in the previous chapter). To imitate this in a lab, we needed a set of formal rules, of which violations would result in collective sanctions. For this purpose, the standard design has been adapted. Participants had to invest a certain minimum into a group project. When they failed to do so (and were detected by a random checking mechanism), the entire group suffered decreased payoffs. Many theoretical arguments defending collective sanctions suggest that the sanctions serve as a delegation mechanism, which would boost the willingness of other members of the group to provide such a second-order public good as peer punishment for non-cooperation. Despite these expectations, the investments to common project when individuals were able to punish their peers under collective sanctions were scarcer than under individual sanctions. One of the striking results of this paper is that collective sanctions are perceived

as deeply unfair (although there is a significant gender difference both in behavior and perception). Thus, we can speculate that collective sanctions fail due to the lack of legitimacy of such a regime: after all, people do not understand why they should be punished monetarily if someone else has failed to meet the minimum investment requirement.

But what if we introduce the ability to punish collectively members of another group for an individual defection? This way of treating outsiders is much more acknowledgeable because all of us have faced prejudice and stereotyping. What effect the endogenous collective sanctions may have on intergroup relations is a topic for the second chapter of this thesis.

The main issue with the intergroup cooperation is that most of the theoretical work points out that we should observe eminent intergroup conflict, whereas in the real life, most groups co-habit peacefully and successfully cooperate (Fearon and Laitin, 1996). This puzzle has been rarely considered both empirically and theoretically. The seminal work by Fearon and Laitin develops the functionalist explanation of the potential mechanism for how intergroup cooperation appears. Collective sanctions are implicitly one of the keystones in their system. According to their logic, the threat of collective sanctions from an outside group makes ingroup policing more effective. The introduction of such an institute also solves several issues of norm enforcement among the outsiders, specifically the high cost of detection and persecution. The higher chance for cooperation due to higher norm enforcement then serves as a self-reinforcing mechanism; contact theory claims that the experience of successful cooperation with the outgroup generates the intention to cooperate in the future and decreases ingroup bias and outgroup prejudice (Pettigrew and Tropp, 2006).

It must be stressed that collective sanctions is just one of many ways to resolve an intergroup conflict. There are numerous other mechanisms that stabilize and appease the intergroup relations such as contact hypothesis (Allport, Clark, and Pettigrew, 1954) or the idea that multiple loyalties that cross cut social circles undermine the motivation for intergroup conflict (Simmel, 2010). The general idea behind all these mechanisms is the *reconceptualization of group categories* (Schiappa, Gregg, and Hewes, 2005), or their *decategorization* (Brewer, 1999). All these approaches can be a way more efficient than collective sanctions, but before these methods are applied there should be a certain pre-existing demand within groups to cooperate. The conditions for these methods to succeed is a set of norms supporting such cooperation and willingness to implement them (Forsyth, 2010, p.432). That makes the situation with implementation of collective sanctions a bit different from the suggested methods.

What if we are yet in the situation when groups do not meet the requirements posed by Forsyth for recategorization? In this case some methods like collective sanctions can be helpful. Another problem with the approach of blurring down the intergroup boundaries is that as it was proved the presence of intergroup competition increases the intra-group cooperation (Erev, Bornstein, and Galili, 1993) so the downside of such deconflictization would be a decline in overall cooperation rate.

In general, in the theoretical literature there is overwhelming pessimism about the prospective for intergroup cooperation. In 1906, William Sumner claimed that "loyalty to the group, sacrifice for it, hatred and contempt for outsiders, brotherhood within, warlikeness without – all grow together, common products of the same situation" (Sumner, 2013), and recent work has stated that the "the current and dominant view in the social and human sciences (including psychology) is that hostile competition is a main driving force in intergroup behavior" (Stürmer and Snyder, 2009, p.6). There

are two main theoretical explanations for why groups should be in conflict with each other rather than to cooperate peacefully: the rational explanation and the psychological explanation. Both schools of thought follow very different logics but agree on the one point: the chances of intergroup cooperation are slim (Wilson, 2015). The rational line of thinking claims that since adjacent groups have to share the same territory, they compete for the limited resources and the only logical way to survive is to push competitors out or to eradicate them completely (Gellner, 2008). The psychological line of thought, dominated by social identity theory (Tajfel, 1982; Tajfel and Turner, 1979), states that the feeling of belonging to a group has an intrinsic value for individuals. This feeling is created and reinforced through the mechanism of estrangement from other groups, so even without any kind of competition for resources, in-group bias quickly appears. The second chapter tries to empirically test Fearon and Laitin's hypothesis of beneficial effect collective sanctions may have on intergroup cooperation. For this purpose, a group of six players were divided into two equal-size subgroups and the members of each subgroups were matched in pairs with out-group members. They first played a Prisoner's Dilemma game, after which they could observe the decisions of another matched pair and took a decision on third-party punishment. In the Collective Sanctions treatment, the punishment applied to an out-group member was randomly assigned to one out of three members. As we shall see the introduction of collective sanctions decreased the cooperation in Prisoner's Dilemma game and substantially suppressed an out-group punishment, proving that collective sanctions are deleterious for intergroup cooperation despite Fearon and Laitin's expectations.

Chapter 3 analyses how the introduction of collective sanctions affects willingness to punish norm violations. In this lab experiment, people were randomly matched in groups of two and had to make decisions first in a binary Dictator game towards a charity. If they chose to take the money from

the charity to increase their personal payoff, they were fined with a certain probability. After that stage, they were also able to punish their peers for making norm-violating choices and this punishment was costly. In the collective sanctions treatment, norm violation could result in a decreased personal payoff for both members of a group. The results demonstrated that collective sanctions significantly increased the frequency of peer punishment, but this did not serve as an effective deterrence mechanism enforcing the norm. There was no significant difference in the amount of money withdrawn from the charity between treatments. The difference in punishment behavior was driven by the behavior of those who chose to violate the norms themselves: while only 12% of the norm violators in individual sanctions punished their norm-violating peers, the amount of those under collective sanctions grew up to almost one third. This confirms the intuition that collective sanctions increase the amount of 'spiteful' punishment in a group because it provides a punisher with an extra motivation. The collective sanctions also served as a mechanism undermining social relations in a group: the unwillingness to punish their peers "without substantial reasons" dropped when the regime changes from individual to collective sanctions, while the fear of retaliation increased.

Chapter 1

Strike one hundred to educate one: Can collective sanctions be efficient?

Abstract

In this paper, we test if sanctions applied to an entire group for the free-riding of one of its members can increase the level of cooperation within that group. To measure the efficiency of such collective sanctions, we conducted a lab experiment based on a standard public good game. The results show that overall, collective sanctions are inefficient. Moreover, when subjects are able to punish their peers, the level of cooperation is lower in the regime of collective sanctions than under individual sanctions. Both outcomes can be explained by a general disapproval of the collective responsibility for an individual fault: in the post-experimental survey, an absolute majority evaluated such regimes as unfair. But although collective sanctions are not an effective means to boost group compliance, there are nevertheless two insights to be gained here. First, there are differences across genders. Under collective sanctions, males' level of compliance is substantially higher than under individual sanctions while the opposite is true for females. Second, there were intriguing differences in outcomes between the different regime types. Under collective sanctions, a person who is caught tends to comply in the future, at least in the short term. In contrast, under individual sanctions, an individual "wrongdoer" decreases his or her level of compliance in the next period.

Keywords: Collective sanctions, Public good game, Crime deterrence

1.1 Introduction

In 2014 the Narxoz University in Almaty, Kazakhastan hired a new rector, Polish economist Krzysztof Rybinski. His main task was to clean up the overpowering corruption and fraudulent practices pervasive in the university. Despite knowing that in the majority of post-Soviet universities there was a wide-spread and severe lack of integrity, Rybinski was still shocked by the level of corruption, cheating, and plagiarism that he encountered.

Among other bribery curbing measures, Rybinski introduced collective sanctions: if any employee was caught behaving corruptly, he or she, plus their superior were fired. As the rector explained, since instituting these collective sanctions, there had not been another single case of bribery among teaching staff. When asked whether this approach provided an incentive for managers to provide cover ups for their subordinates, Rybinski answered that if a manager reported the misdeed, he or she was protected from the punishment, meaning there was no incentive to lie on behalf of employees (Matthews, 2016).

Even if collective sanctions may be an efficient measure to prevent certain kinds of criminal actions, their usage goes against the entire logic of modern justice, which is based on the idea of retribution. As Kant states, "judicial punishment can never be used merely as a means to promote some other good for the criminal himself or for civil society, but instead it must in all cases be imposed on him only on the ground that he has committed a crime" (cited by Ross 1975, p.54).[1]

Advocates of collective sanctions usually justify them with two different lines of argumentation (Peterson, 2012, p.169). First, the other group members are guilty of negligence. They had a chance to prevent an antisocial

[1]For further philosophical discussion of collective sanctions and punishment see Corlett 1992.

action by a team member, but they preferred to stay idle. So collective sanctions are, in reality, individual sanctions for omission. This is an attempt to solve the conflict between an idea of punishing the collective and Kantian idea of retributive justice, which assumes only individual responsibility for a criminal action. Since idleness in correcting a team member's behavior is treated as antisocial action itself, collective sanctions are intended to correct inaction and to increase the amount of peer control.

The second argument follows a radical consequentialism, or a rational cost-benefit analysis: it does not matter that group members are not directly guilty in antisocial actions of the specific member but punishing collectively is warranted on the grounds of efficiency. As Levinson, 2003 stated in his overview of collective sanctions: "*Group members might be punished not because they are deemed collectively responsible for wrongdoing but simply because they are in an advantageous position to identify, monitor, and control responsible individuals, and can be motivated by the threat of sanctions to do so.*" (p. 348). This logic is built upon the idea of delegation of responsibility by an outside authority. If the entire group is punished for the misdeed of one member, it becomes the individual members' task to detect and prevent antisocial behavior. The positive consequences of delegating the responsibility to detect and prevent crime on the nearest neighbors of a perpetrator outweigh the harm brought by punishing an innocent.

In both cases the conclusion is the same: the introduction of collective sanctions converts the task for an outside authority to find a wrongdoer (a free-rider in case of a public good setting) into the task of his/her peers to detect, prevent and punish the perpetrator. This paper examines what kind of consequences collective sanctions have on cooperation within a group, and on the willingness to punish uncooperative behavior by peers. Thus, the main objective of this paper is to answer the following question: **Can collective sanctions for an individual's antisocial behavior be beneficial for**

cooperation?

The paper is organized as follows. In Section 2, I review the current lay of the land in the study of collective sanctions. In Section 3, I describe my experimental design. In Section 4, I present results and analysis from the experiment. Finally, in Section 5 I conclude and offer suggestions for further inquiry.

1.2 Theoretical considerations and empirical evidence

By definition, collective sanctions (CS) are imposed on an entire group for a crime or misbehavior committed by a single member of that group. Pereira et al., 2015 define CS as "the negative treatment inflicted by authorities or by an outgroup upon an entire social group, in reaction to an offense committed by one or some of its members". It is usually traced back to pre-modern or primitive societies where it was a key concept of law. Berbers for instance project an idea of 'guilt' for murder upon the nearest relatives of the murderer: "if a homicide takes place, the ten closest agnates of the offender are immediately at risk because they are equally 'culpable'"(Gellner, 2000, p.376). It is possible to erroneously come to conclusions that in modern life collective sanctions are limited mostly to military bootcamps and prisons (cases mentioned by Heckathorn 1989 and Whitmeyer 2002 in their theoretical works on CS). But in fact, many policy makers across the globe are proponents of this measure.

Policy makers face the problem that resources to enforce the law are limited. The logic of CS is to delegate the power to detect norm violators downwards to the group members, and to grant them the authority to deal with a wrongdoer themselves.

For instance, after the democratic revolution in 2014, the new Ukrainian government had to deal with the flagrant corruption in the customs service. On May 17th, 2016 the newly appointed head of Customs Service, Hennadi Moskal, wrote on his Facebook page: "Today, the shift foreman of customs at the Ukrainian-Slovak border was caught by us in a bribery…For 420 Euros, he promised to let some smugglers with cigarettes go. Last week at a meeting with all the heads of the customs offices I promised: if anyone will be caught accepting bribes, I will discharge the entire shift. So in our case, for 420 Euros, one bribe-taker ruined his career, and the career of all his subordinates. We will do the same with other bribe takers."[2]

Another area in which collective sanctions are widely used is the school system. Teachers often give a detention to an entire class when only a few misbehave. One recent and eyebrow-raising example of a collective sanction happened in 2010 in a Manhattan school, where the principal banned all 2,445 students from using toilets for the entire day because of a fight that happened one day earlier (Schuck, 2010).

The belief that collective sanctions can be successful in curbing norm violations is common not only among policy makers, but also among academics. For instance, in a review of solutions to collective action dilemmas, CS are listed as a tool to boost informal control in a group: "A common control technique is to punish the whole group for some act committed by one of its members. If the punishment is severe, as it often is, this technique may be horrendously effective" (Monahan and Walker, 2011).

But if CS are widely used, or at least recommended by policy makers, do they produce those beneficial effects on individual behavior? And why do they do so?

There are some theoretical suggestions that collective sanctions, under certain conditions, indeed can increase individuals' willingness to cooperate,

[2]https://www.facebook.com/hennadii.moskal/posts/922829371195440

or, vice versa, deter people from free-riding and non-cooperation. Following Nakao and Chai, 2011 work on collective punishment, these arguments can be divided into functional, preferential, and informational arguments.

First, the functional argument, central to the logic of CS, claims that the introduction of collective sanctions increases the efficiency and willingness of other group members to conduct ingroup policing.

Some authors consider the internal control capacity affected by collective sanctions as the only factor that can make CS effective. One strand of research is rooted in the rational choice literature and cites pragmatic reasons for group members to react to CS. In Heckathorn's model, "group members have incentives to urge one another to seek out external sources of rewards and to comply with external dictates to avoid triggering externally induced punishments" (Heckathorn, 1990, p.367). Another argument is rooted in social identity theory, which explains cooperation and norm compliance by the commitment of an individual to the group he feels he belongs to (Tajfel, 1982). People tend to cooperate more with their own group members and the costly punishment of group members for norm violation is itself a second-order public good. If a person strongly associates him/herself with the group, that increases the so-called black sheep effect, which is the tendency to punish their own group members more severely than outsiders (Marques, Yzerbyt, and Leyens, 1988; Shinada, Yamagishi, and Ohmura, 2004). Collective sanctions, by producing a common experience for the group, increase group cohesion, resulting in a larger 'black sheep effect' and increased propensity to punish norm violators.

Second, collective sanctions may work because they change the **preferences** of a wrongdoer him/herself. The introduction of CS can be more effective than individual sanctions, because of the additional punishment brought upon the other members of the group. The total 'amount' of sanctions that may be assigned is thus higher under CS. If a wrongdoer cares about the

suffering of others, the prospect of causing them to be punished may deter him or her to engage in the devious action in the first place. For example, the tradition of mutual exchange of hostages in some tribes guarantees that a member fulfills the contract, otherwise his kin can be harmed.

Third, the **informational** argument states that for an external authority, it is hard to detect who is guilty of an antisocial act, but for ingroup members, this task is relatively cheap and attainable. So collective sanctions will increase the detection rate by group members, the argument goes, while the punishment itself can still be carried out by an external group. Thus, collective sanctions address the information asymmetry that exists between ingroup and outgroup members.

Where does the field stand in terms of empirical research on collective sanctions? Richard Posner starts his essay on the economics of collective sanctions by describing them as "a conventional legal tool that is efficient in many of its applications" (Becker and Posner, 2009). To-date, this and related claims have not yet been thoroughly tested empirically: to the best of my knowledge, there have only been three lab experiments on collective sanctions. In an unpublished paper, Dickson (2007, p.5) claims that his study "appears to be the first lab experiment involving collective punishment", and in their paper on random sanctioning, Fatas et al. claim that, "*As far as we know, no experimental analysis of random punishment in teams has ever been done*" (Fatas, Morales, and Ubeda, 2010, p.360).

Dickson's 2007 study is similar to the one presented in this chapter. In his paper, participants engaged in a standard public good game, where players chose how much money to invest in a group project. One of five group members was randomly assigned the role of "central authority" and was able to punish other group members collectively. In one of two treatments, the interest of the central authority was aligned with the interest of the group: his or her earnings increase with the amount invested in a group project. In the

other treatment, the interests of the group and that of the enforcer opposed one another. Dickson found that collective sanctions had a subtle, short-lived positive effect on cooperation in the case of aligned interests, and a strictly negative effect in the case of opposed interests. However, the design also suffered from some weaknesses, namely the fact that the principal was part of the group, that the cost of punishment varied across treatments, and that there was a 100% probability of detection. Arguably because of these weaknesses,punishers tried to avoid using collective sanctions, leading to an overall very low frequency of punishment.

Fatas, Morales, and Ubeda, 2010 provide an additional angle on collective-sanctions-testing in the lab. In their experiment, a randomly chosen member is punished by exclusion from the group (and from getting his/her share) if groups are found to have low contributions. The participants found this approach procedurally unfair, but it significantly boosted cooperation. The authors do not treat their experiment as a study of collective sanctions, but it is plausible to interpret it as such, as Levinson, 2003 does, writing: "[i]n general, so long as groups are sufficiently solidary, group incentives will be the same whether collective sanctions are lumped on one member of the group chosen at random (or by any other criteria besides culpability) or spread evenly among all group members" (p. 377). However, in Fata, Morales and Ubeda's 2010 design, the probability of exclusion grew linearly with the number of violators, which made it rational for participants to cooperate when the expected frequency of violations changed. In this sense, the efficiency of collective sanctions was not tested by this design but was rather implied by construction. Thus, their design does not allow us to disentangle the effect of collective liability in and by itself on the efficiency of cooperation.

The most recent study by Yefeng, Jiang, and Villeval, 2015 focuses on the effect of collective sanctions on corruption levels among bureaucrats. They

used the same design as Abbink, Irlenbusch, and Renner, 2002, with one important amendment: If the number of corrupt bureaucrats exceeded a certain threshold (in their case, there were two treatments with thresholds of 20% and 60%), this triggered collective failure, which meant collective sanctions and decreased payoffs for the entire group of bureaucrats. This threat was not particularly effective; the risk of collective sanctions did not avert the firms from offering a bribe and the number of officials who chose to accept it (75%) was much higher than predicted.

The main aim of this paper is to test whether the introduction of collective sanctions result in higher degree of cooperation, and how collective sanction interact with peer punishment (that is, with the functional argument).

I design an experiment with a 2×2 design. The experiment crosses the institutional regime (individual vs. collective sanctions for a failure to invest enough into a public good), with the presence or absence of ingroup policing.

1.3 Hypotheses

Collective sanctions affect the decision of an individual regarding free-riding in the production of a public good in two different ways: directly and indirectly. CS change the individual preferences **directly** by making norm violations costlier.

Norm violators incur a psychological cost (Abeler, Becker, and Falk, 2014; Cohn, Maréchal, and Noll, 2015). As Nakao and Chai noticed, the knowledge that someone else from the group will be punished for free-riding increases the moral costs of such an action (Nakao and Chai, 2011).

On the other hand, when a collective sanction harms a cooperative person, although that person did not actually free-ride, this produces a de-motivating signal, and can result in unwillingness to cooperate later. The punishment of a cooperator can be interpreted as an antisocial punishment (even if it

was not intentionally so) (Herrmann, Thöni, and Gächter, 2008) and there is plenty of evidence that such kind of punishment has a significant detrimental effect on cooperation, both when intentional (Fatas and Mateu, 2015), or when generated by a noisy environment that does not allow the punisher to correctly identify a free-rider (Grechenig, Nicklisch, and Thoeni, 2010).

The overall direct effect thus is unknown and depends on the degree of group cohesion and the probability of being punished for the actions of others, which, in turn, depends on the size of the group. Group cohesion has a beneficial effect on cooperation because it assists a third-party monitoring of norm violations and norm enforcement (Agrawal and Goyal, 2001). The effect of group size in case of collective sanctions is more complicated. As Heckathorn has shown in the case of the Prisoner's Dilemma, the effectiveness of collective sanctions grows with group size up to a certain point, and then starts to decline because the chances to detect and prevent the crime inside a large group becomes smaller for an individual group member, while the chance of being punished grows (Heckathorn, 1988).

The **indirect** effect of collective sanctions is a result of delegation and increased ingroup policing. When possible, that is, CS should induce group members to police each other, which will drive up cooperation rates. Therefore, the institutional regime in which CS are applied should matter. The institutional choice whether CS are applied or not should thus interact with another institutional choice: whether peers are allowed to punish free-riders or not. The indirect effect is believed to be a main mechanism that can explain why collective sanctions should deter crime. It adjusts the information disparity that an external authority has with regard to the perpetrator ((Becker and Posner, 2009, p.303)), and makes people more inclined to deter their own group members from injurious behavior (Becker and Posner, 2009, p.307). In non-criminal cases, for example in the credit markets with third-party liability, such as the Grameen bank program, each member in a group of people

serves as a co-guarantor for everyone else in that group. That guarantees mutual monitoring: "agents influence the other agents' costs of engaging in desirable and undesirable aspects" (Varian, 1990, p.155).

These considerations define the following hypotheses to be tested empirically in the lab:

1. When peers are able to punish free-riders within their group, they will do it more frequently and to a greater extent under the threat of collective sanctions rather than when there is merely a threat of individual sanctions (IS).

2. Due to the expected larger extent of peer punishment under the regime with collective sanctions and the ability of peers to punish free riders, the level of cooperation will be higher here than in a similar regime with individual sanctions (IS) only.

Under a regime of collective sanctions without peer punishment, there are two concurrent processes going on: (1) the moral cost of free-riding grows, encouraging cooperation, and (2) cooperators get punished, and thus are demotivated from further cooperation. So, there are two alternative hypotheses to test under collective sanctions (CS) without ingroup policing, which we will compare to a regime with individual sanctions (IS). Depending on which of these two effects prevail, cooperation may either increase or decrease, and may be higher or lower in CS than in IS:

3a. The level of cooperation under CS will be higher compared to the level under IS because of the higher moral costs of free-riding.

3b. The CS regime sends mixed signals towards cooperators because they can be punished while cooperating. This mixed signal can reduce their willingness to cooperate in the future. So, under CS we shall observe a lower level of cooperation than under IS.

1.4 Experimental design

The basic framework of this experiment was a classic standard public good game, which was played with and without peer sanctions, and with or without the possibility of collective sanctions. This design is represented in the 2×2 Table 1.1 along with the notations assigned to each treatment. I programmed all procedures using zTree (Fischbacher, 2007a).

TABLE 1.1: 2×2 matrix of treatments

	No Peer Sanctions	Peer Sanctions
Individual Sanctions	CP0IS0	CP0IS1
Collective sanctions	CP1IS0	CP1IS1

The experiment consisted of 15 periods; treatments with peer sanctions had three stages per period, and treatments without peer sanctions had two stages per period. Participants were divided into groups of three and were provided with an endowment of 20 tokens each. In Stage 1, they took decisions how much to investment into a group project. In Stage 2, an external check of individual contributions was performed; in Stage 3, participants could use deduction tokens for peer sanctions.

The group composition remained fixed across all 15 rounds (*partner matching*), but the identities of specific participants in a group were not revealed to avoid retaliative strategic punishment or non-cooperation between rounds.

The first stage consisted of a standard public good game where individuals face the choice of whether to cooperate or free-ride. This part was the same for all four treatments, but the anticipation of possible consequences at later stages may convince a person to contribute more or less at this stage, depending on the institutional regime (CS or IS), and potential peer sanctions at Stage 3.

Stage 2 is the only stage where CS and IS treatments differed. Each group's contributions were checked with the same probability (1/3, more details are provided below), but the consequences were different. In the case of individual sanctions, if a person did not meet a minimum threshold requirement, and the external check revealed this, he or she bore individual consequences. In contrast, in the case of collective sanctions the entire group's payoffs was reduced. The last, third stage appeared only in treatments with peer punishment.

The utility function for the first stage coincided with Fehr and Gachter's stranger-treatment public good game (SPGG)

$$\pi_i = y - g_i + a \sum_{j=1}^{n} g_j$$

where y is the initial endowment, g_i is the contribution to a group project, $n = 3$, and $0 < a < 1 < na$ is the return on group project investment. An investment multiplier of 0.5 was chosen, as this coefficient makes calculations easier for participants.

To introduce an element of external authority that imposes collective or individual sanctions on a group with a certain probability, I employed an automatic mechanism that periodically controlled whether individual contributions met a certain threshold. This threshold was set at half of the total endowment: out of 20 tokens of endowment, 11 'should' be invested into the group project. This prescribed number was not presented as the participants' duty, and no morally loaded words (e.g. 'authority' or 'punishment') appeared in the instructions. Instead, the participants were informed that, with a certain probability, their contributions would be checked; if contributions were found to be less than a set threshold, their earnings for that particular period would be diminished – the exact text explaining this varied according to the specific experimental treatment.

The randomized checks were implemented as follows. I uploaded a matrix of pre-generated random numbers from 1 to 100 onto the zTree server. Each group had an associated vector of 15 random numbers drawn from this matrix, one random number per period. In each period, if a number associated with this group and this period was less than 33, then the contributions of an entire group were checked. Thus, the probability that a given group's contributions were checked was 1/3 in each period. This mechanism was explained to participants in a simplified manner to make the explanations clearer while avoiding participant deception. For example:

"In the second stage, there is a 33% chance that the contributions of everyone in your group are checked by a computer. Specifically, during every period, the computer generates a random number between 1 and 100 for each group. If the generated number equals or is lower than 33, then it checks the contributions of all group members in that group."

By pre-generating the numbers instead of generating them inside the experiment, I was able to guarantee that in each treatment there were the groups with the similar history of external controls. Since the order and frequency of external checks influence the intentions of individuals to cooperate and punish the peers in subsequent periods, this design allowed me to control for a history of 'checks' in each of our four treatments.

The different regimes of sanctions were implemented in the following manner. In the individual sanctions (IS) regime, if an individual's contribution was found to be 10 tokens or less during the check, `that participant's earnings` for that period were reduced by 7 tokens. If the group's contributions were not checked, then all individual earnings just took home the amounts they earned during that round.

Under the collective sanctions (CS) regime, if the contribution of at least one group member was found to be 10 tokens or less during the check, the earnings of everyone in the group in that period were reduced by 7 tokens.

Therefore, after the second stage of checking, the profits of individual members were

$$IP = \begin{cases} \pi_i - F & g_i < T \cap p < \frac{1}{3} \\ \pi_i & g_i > T \cup p \geq \frac{1}{3} \end{cases}$$

$$CP = \begin{cases} \pi_i - F & g_j < T \cap p < \frac{1}{3} \\ \pi_i & g_j > T \cup p \geq \frac{1}{3} \end{cases}$$

Where $j \in 1 \ldots n$, CP is the profit under collective sanctions, IP is the profit under individual sanctions, $F = 7$ is the fine if a contribution is less than the threshold $T = 11$, and p is a random number between 0 and 1.

At Stage 3, in treatments with peer sanctions, participants were able to send deduction points to the other members of their group, up to a maximum of 10 deduction points for each of the peers. Each deduction token reduced the recipient's earnings by 2 tokens, while reducing a sender's earnings by 1 token.

1.5 Game-Theoretical predictions

The expected amount of the fine imposed by a central authority if a subject fails to invest above the necessary threshold of 10 tokens is calculated as the probability of being caught (p), multiplied by the amount of the fine (F). A net loss of investment of the threshold T is $(1 - a)T$, where a is the rate of return on investments to a common pool. So unless $pF > (1 - a)T$, a rational profit-maximizer will behave in the same way s/he would behave in a regime without a contribution threshold. The same logic applies in the case when costly peer sanctions are introduced. These peer sanctions are a second-order public good, so there is an incentive to free-ride in their production. The purely game-theoretical (but not behavioral!) prediction therefore

is that people do not make use of peer sanctions (As it is known from Fehr and Gächter, 2002b or Herrmann, Thöni, and Gächter, 2008 people punish their peers ignoring the rational profit-maximizing considerations).

Thus, under an individual sanctions regime, on average the same equilibrium should be observed as in any other standard public good game with peer punishment stage – no matter what kind of preferences the participants have towards the peer sanctions: if participants expect that non-cooperative behavior is punished by the peers, we shall observe the convergence towards full cooperation, or if people fail to provide this second-order public good, the cooperation will decline. When collective sanctions are applied an optimal strategy depends on the size of j, an expected number of violators. Even if the probability of a group being checked is the same as it was under individual sanctions, the chances to be sanctioned externally grow with the expected number of wrongdoers. That is the reason why the efficiency of collective sanctions drop with the growth of a group size Heckathorn, 1988: as the group gets larger, so does the chances to be externally sanctioned. In groups of a significant size under collective sanctions norm compliance is not a viable strategy to avoid sanctions. The burden of being in such a group is additionally worsened because an individual participant is in a less advantageous position from the point of view of information: he or she may not know who was an actual perpetrator so he feels helplessness, being punished by an external force without being able to detect a norm violator who was a cause of these sanctions. On the other hand, in smaller groups the introduction of collective sanctions increases the probability of peer punishment thus, we can expect the growth of norm compliance. Since the vectors of these two mechanisms (lower cooperation rate in expectation of being punished even you cooperate and higher expectation rate due to expected peer punishment) are oppositely directed, without specific parameters (such as a group size and an expected frequency of norm violation) it is hard to give

clear-cut theoretical predictions whether the equilibrium would differ from an individual sanctions regime.

It is important to detail how we chose these exact parameters for a game. The key factors that defined the game were p, F, and the cost of deduction tokens c_d. Since Fehr and Gächter, 2000 work on peer punishment, various studies on SPGG have used $c_d = 3$. As Carpenter, 2007b found, the demand for punishment is inelastic and approaches zero for any price below 1. At the same time, as Nikiforakis and Normann, 2008 analysis of the statics of punishment in SPGG showed, for $c_d = 3$, the level of cooperation reaches almost 100% after approximately the 5th round, 0% for $c_d \leq 1$ (confirming Carpenters findings), and remains stable at around 50% for $c_d = 2$. If we would like to see an effect (in whatever direction) of collective liability on contributions, then $c_d = 2$ appears to be a safe option because, as Nikiforakis and Normann's paper shows, that is the price of punishment which makes roughly half of the population to cooperate, thus resulting in the highest variability. Finally, in regard to the group size, while $n = 3$ may be an unusual size (SPGGs follow Fehr and Gachter, who chose $n = 4$), both lab studies (Carpenter, 2007a) and a meta-analysis of 27 SPGGs (Zelmer, 2003) did not find a significant effect of the group size on contributions or punishment levels.

1.6 Results

The experiment was conducted in the Columbia Experimental Laboratory in the Social Sciences (CELSS) using the standard z-Tree (Fischbacher, 2007b) software. The design was approved by the Columbia University internal review board, participants were recruited via the ORSEE online system. Before proceeding with the experiment, all participants signed a consent form according to the IRB protocol. Subjects received a guarantee that their decisions as well as their payoffs would remain completely anonymous. In total, 108

participants took part in the experiment during 8 sessions, which took place between December 1st and 4th of 2015. The number of participants in each of the four treatment groups is shown in Table 1.2. Instructions are available in Appendix A, and screenshots of the z-Tree program are shown in Appendix B.

TABLE 1.2: Number of participants per treatment

Treatment	Peer punishment	Collective sanctions	Participants	Observations
CP0IS0	No	No	24	360
CP0IS1	Yes	No	30	450
CP1IS0	No	Yes	24	405
CP1IS1	Yes	Yes	27	405
Total			**108**	**1620**

The average payment the participants received at the end of the experiment was $22.00 (all currencies are US dollars), including $5.00 as a reward for showing up. Earnings varied between treatments, being slightly higher for individual sanctions ($22.40 vs. $21.70 in CS) and for treatments without peer sanctions ($22.20 vs. $21.90), but statistically the difference was not significant.

The dynamics of individual contributions into a group project show similar patterns for the collective and individual sanctions regimes (Figure 1.1). All participants started with high contribution levels of 10 to 12 tokens out of 20. Without peer sanctions, cooperation began to decline after the 5th or 6th round to contributions of 5 or 6 tokens out of 20. This pattern echoes observations in other voluntary contribution experiments (see contributions in baseline scenarios with no punishment in such studies as (Bochet, Page, and Putterman, 2006; Page, Putterman, and Unel, 2005; Cason and Khan, 1999), and is considered a common phenomenon explainable by strategic behavior or learning (Andreoni, 1988).

With peer sanctions, the average contributions remained relatively stable at about half of the endowment (10-12 tokens) until the 15th (and the last) round, when the contributions dropped – again, a typical effect of the 'end game' for other Voluntary Contribution Mechanisms (VCM) with sanctions. When peer sanctions were available, CS contributions were lower than IS. There was no such difference in CS and IS treatments without peer sanctions.

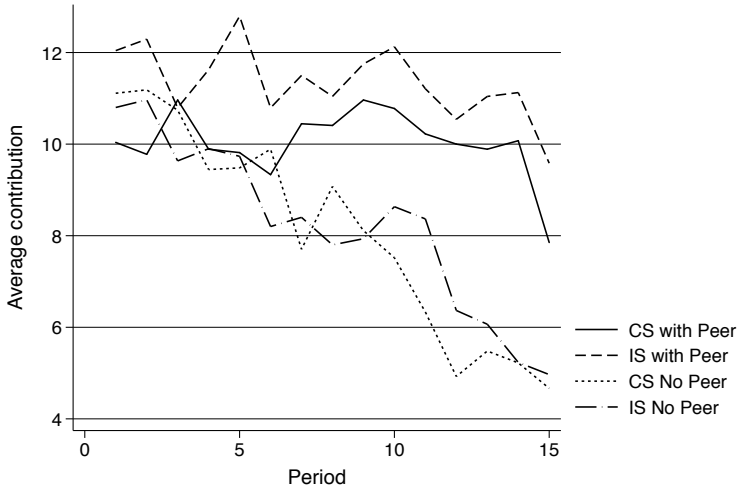


FIGURE 1.1: Average contribution to the group account in each period, by treatment

The subjects could choose to invest any number of tokens (between 0 the total endowment of 20) into a group project, with the safe threshold of 11 tokens, below which an external punishment could be applied. In reality, their choice set was much more limited. 81% of contributions fell into one of three categories:

1. 31% (502 observations) of contributions were 0, or total non-compliance;
2. 28% (456 observations) of contributions were 11, exactly 'at the edge' of compliance;
3. 22% (361 observations) of contributions were full cooperation of 20 tokens.

This trimodal distribution provides an additional layer of analysis. When rules define a threshold for bare minimum cooperation, a rule-follower has a choice to be a marginal cooperator who contributes right above the necessary threshold, or to voluntarily cooperate to a degree larger than required.

The patterns of cooperation/non-compliance with regard to a threshold vary across the treatments (Figure 1.2). In general, CS again proves its ineffectiveness: the share of pure non-compliers (those who contribute less than a threshold, $gi < T$) is higher under collective sanctions than under individual sanctions. That is true for treatments both with and without peer sanctions. Without peer sanctions, the percentage of non-compliers under CS is 51% vs. 47% under IS, and with peer sanctions non-compliers represent 36% under CS vs. 31% under IS.

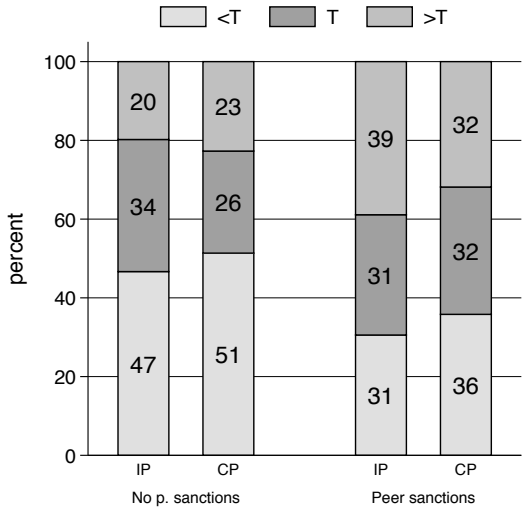


FIGURE 1.2: Percentage of contributions below and above threshold

When we focus on those contributions which met the requirements of 11 tokens or more only (that is, those who were either at the edge of compliance or who were full voluntary cooperators), we observe diverging behavior between treatments with and without peer sanctions (Figure 1.3). Without the

possibility of being punished by group members, the share of 'bare compliers' is lower under collective sanctions (only 53% of all compliers). The behavior changes when we introduce peer sanctions; the number of voluntary cooperators is now higher (56% vs. 50%) in the individual sanctions regime. Presumably, peer pressure makes people more reluctant to break the minimal-contribution rule, but at the same time that "kills" their motivation to fully cooperate above the mandatory level.

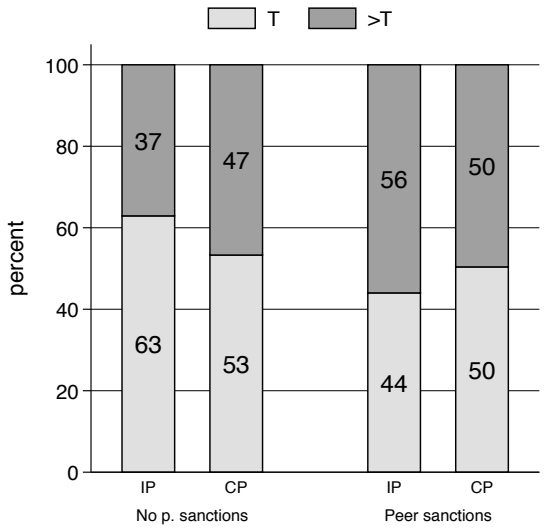


FIGURE 1.3: Percentage of contributions above threshold only

Additionally, there were unexpected, but clearly visible, differences in behavior between genders across treatments with peer sanctions (Figure 1.4). There is currently no consensus in the literature about gender differences in contribution levels in public good games. Some studies have found no gender difference in contributions (Sell, Griffith, and Wilson, 1993; Andreoni and Petrie, 2008), some have found that women contribute more (Seguino, Stevens, and Lutz, 1996), and others have found that males contribute more (Solow and Kirkwood, 2002). Still others find more nuanced effects, such

as the observation that women contribute significantly more when the free-riding option is intentionally framed as a harm to the rest of the group (Fujimoto and Park, 2010), or that women start with higher levels of contributions – and effect that fades out over time, however (Cadsby and Maynes, 1998). Balliet et al.'s 2011 large meta-review of more than 270 studies of social dilemmas points out that generally, there are no significant gender differences, although men tend to be more cooperative in repeated interactions.

The two dominant theoretical frameworks in this area also have contradicting predictions. First, the sociocultural theory explains the difference in behavior across genders as differences in social roles and experiences that individuals sustain throughout their life course (Eagly and Wood, 1999; Eagly and Wood, 2011). Since women, due to their social positions, develop more interpersonal skills, they are more community-oriented and thus contribute more and free-ride less in social dilemmas. In contrast, the evolutionary paradigm assumes that the gender differences evolved as a result of adaption to different problems in the course of evolution, namely mostly hunting for men, and gathering for women (Hawkes and Bliege, 2002; Silverman and Eals, 1992). According to this line of thought, since hunting needs more group coordination than gathering, men cooperate in social dilemmas more than women.

These contradictory theoretical and empirical stances influenced the decision not to include gender differences in behavior under collective sanctions regimes in the guiding hypotheses. In the course of analyzing the results, though, important empirical contributions to this larger area of inquiry emerged, and are thus included in the analysis and results section of this paper.

It turns out that male subjects contribute less than female subjects in IS (on average 10 tokens vs. 12 for females) and significantly more in CS (14 vs. 8, or +75%). No such pattern is observed in treatments without peer sanctions

(Figure 1.5). The same is true if we look only at the contributions above the contribution-threshold. On average, under IS, the female participants who decided to 'obey the rules' invested 16 tokens, but invested only 12.8 tokens under a CS regime. The situation is exactly opposite for males (13.8 under IS vs. 17.0 under CS). The proportion of voluntary cooperators among females in IS is 60%, but almost three times (22%) as low among males (see left plot of Figure 1.6). The situation is the opposite under collective sanctions, where 64% women are "bare" contributors, compared to only 29% of males.

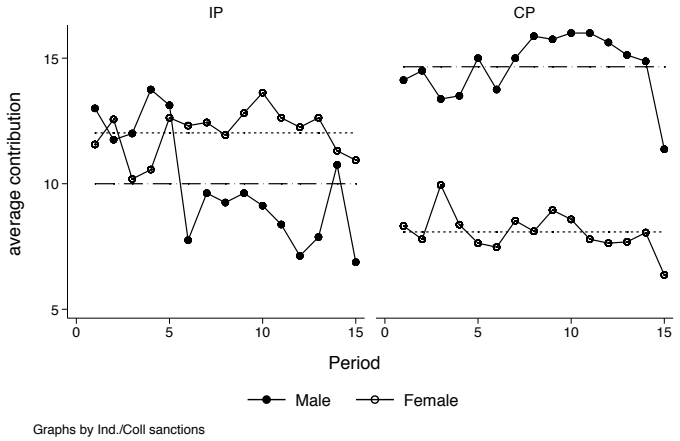


FIGURE 1.4: Average contribution by gender in treatments with peer sanctions

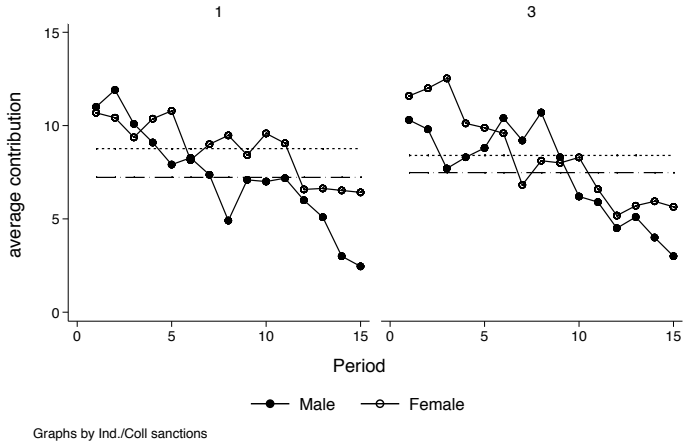


FIGURE 1.5: Average contribution by gender in treatments without peer sanctions

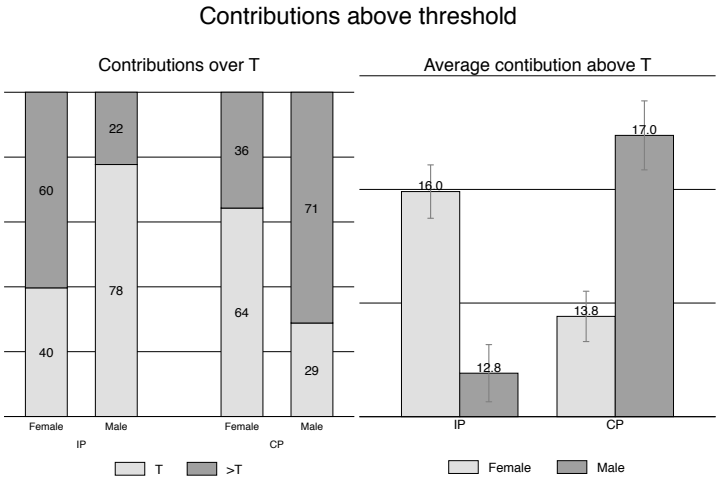


FIGURE 1.6: Contributions above threshold across gender

A cross-sectional OLS (Figure 1.7) where the dependent variable was the average contribution of an individual across all 15 periods confirmed that the difference across genders is statistically significant at the 10% level for the treatment with peer sanctions. Males under collective sanctions contributed significantly more than females, even controlling for their estimations of fairness of the specific regime. As we shall see later, the fairness evaluations vary substantially across genders, and females find collective sanctions much more unfair, so the overall gender effect is even larger.

We followed Gaechter and Renner, 2010 and Anderson and Putterman, 2006 for panel data analysis, running a random-effects Tobit regression and random-effects panel OLS (both sets of models are shown in Tables 1.3 and 1.4, respectively). Tobit models are widely used for analysis of voluntary contribution studies (see Solow and Kirkwood 2002) due to the fact that possible contribution levels are bounded from below and above, meaning that ordinary OLS can be biased. The panel data analysis confirmed the cross-sectional OLS results: again, males under the CS regime with peer sanctions demonstrated significantly higher levels of contributions. This is observed in the large and statistically significant effect of the interaction term 'CS X Male'

which approximately twice outweighs the negative effect collective sanctions have on the predicted level of cooperation.

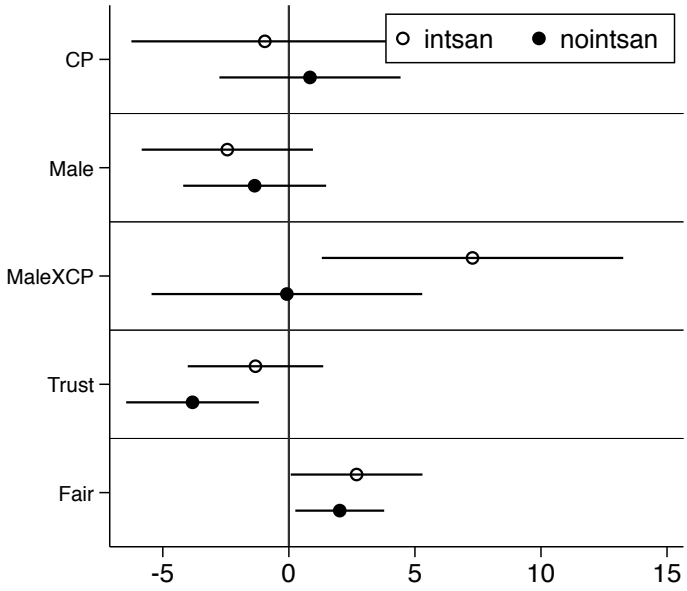


FIGURE 1.7: OLS regression. DV: Contribution, $n = 108$, 10% CI

This gender-based behavioral difference in reaction to collective sanctions can be explained by a gender-based difference in perception of the two regimes. In a post-experimental questionnaire, we asked participants to evaluate the fairness of the specific sanctions rule used in the game. Fairness was estimated by participants twice. First, they graded the regime they experienced in the experiment using a four-level Likert scale (from "very unfair" to "totally fair"). Next, we explained the rules of another treatment (collective sanctions to the participants of the individual sanctions regime, and vice versa). Participants then had to grade the fairness of this alternative regime compared to the one they just experienced. This doubled the number of estimations (with all relevant limitations) and usefully put the evaluation of the regime they had experienced into context.

TABLE 1.3: DV: Contribution to the group project. Tobit random-effect baseline and two extended models

DV: Contribution	**(1)**	**(2)**	**(3)**
Collective sanctions (CS)	-2.090	-6.202**	-6.659**
	(2.457)	(2.891)	(3.115)
Peer sanctions	3.977	4.616**	4.340*
	(2.463)	(2.347)	(2.525)
Gender (male)		-5.236	-6.162*
		(3.440)	(3.692)
CS X Male		11.24**	12.15**
		(4.937)	(5.297)
Trust		-6.642***	-7.405***
		(2.433)	(2.621)
Peer sanctions received$_{t-1}$			0.131
			(0.173)
Peer sanctions sent$_{t-1}$			0.524***
			(0.183)
CS Applied$_{t-1}$			-1.492
			(0.957)
Group is checked$_{t-1}$			2.030***
			(0.701)
Sigma	8.208***	8.209***	7.804***
LL	-3320.37	-3315.13	-2985.74
Wald	3.74	15.04***	32.55***
Observations	1,620	1,620	1,512
Individuals	108	108	108

Standard errors in parentheses

$^{***}p < 0.01,^{**}\ p < 0.05,^{*}\ p < 0.1$

TABLE 1.4: DV: Contribution to the group project. Panel OLS (random-effect) baseline and two extended models

DV: Contribution	**(1)**	**(2)**	**(3)**
Collective sanctions (CS)	-0.698	-2.546*	-2.551***
	(1.172)	(1.402)	(0.965)
Peer sanctions	2.557**	2.858**	2.911***
	(1.174)	(1.139)	(0.786)
Gender (male)		-2.094	-2.358**
		(1.678)	(1.154)
CS X Male		5.032**	5.270***
		(2.393)	(1.646)
Trust		-3.133***	-3.172***
		(1.180)	(0.812)
Peer sanctions received$_{t-1}$			0.0177
			(0.0893)
Peer sanctions sent$_{t-1}$			0.193**
			(0.0826)
CS Applied$_{t-1}$			-1.260***
			(0.472)
Group is checked$_{t-1}$			1.185***
			(0.350)
Sigma	4.579	4.579	4.358
R^2	0.0297	0.0890	0.119
Wald	4.967	15.95	53.82
Observations	1,620	1,620	1,512
Individuals	108	108	108

Standard errors in parentheses

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

While for IS there was almost no difference in the fairness evaluations between males and females, for CS, females found the regime much more unfair; the difference is statistically significant at a 10% level (see Figure 1.8). Perception of the regime fairness appears to be a key factor that explains why CS is not as efficient as it should be. The OLS coefficients of fairness on average contributions are both positive and statistically significant (Figure 1.9), and this relationship is observed across treatments and different estimates of fairness level.

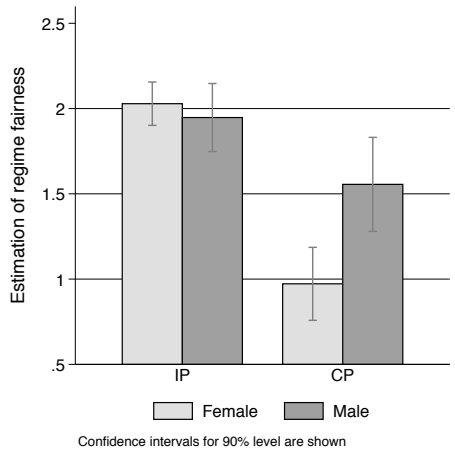


FIGURE 1.8: Fairness estimation by gender in two different regimes

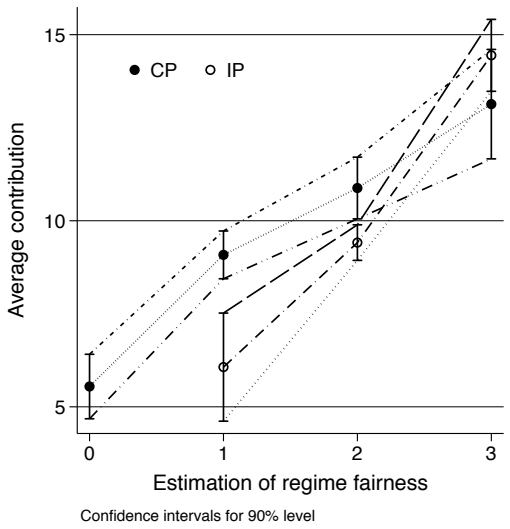


FIGURE 1.9: Contribution levels in different regimes by fairness

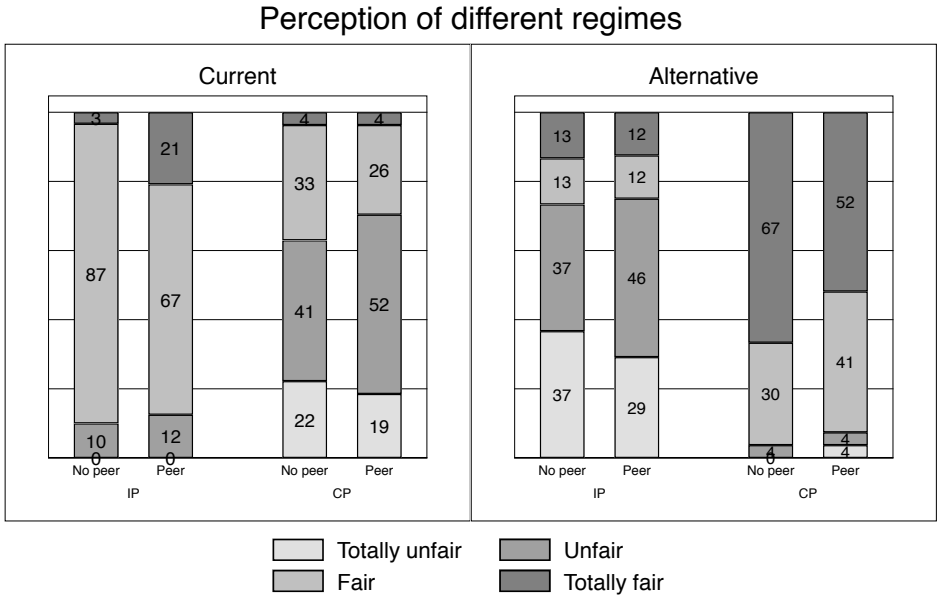


FIGURE 1.10: Regime perception across treatments

1.7 Different reactions to the external check

It could be expected that the previous experience of sanctions by a central authority would affect participants' behavior in the next round. This reaction is observed in most iterated voluntary contribution experiments, like Baldassarri and Grossman, 2011. The overall effect of an external sanctioning regime can be split into two effects: one from being checked, and one from being punished externally (conditional on one's behavior being checked).

Control and punishment by an external sanctioning authority have very different consequences under the collective sanctions (CS) regime as compared to the individual sanctions (IS) regime. Under IS, if the entire group is checked, a person does not bear the external sanctions as long as s/he did not break the rules (in our case, s/he should have contributed more than 10 tokens). Therefore, the external control mechanism can confirm a person's prior beliefs that following the rule is the right decision. However, it may happen that this can provoke the opposite reaction, due to the well-known

gambler's fallacy – individuals' believe that an unlikely event becomes less likely in the future when it has just materialized. This effect would provoke a decrease of compliance during the next period. An illustrative analogy would be ticket control on public transport. A person with a valid ticket, observing a ticket inspector, may either feel reassured that the purchase of the ticket was the right decision, and that decision should be repeated because ticket control is present, or, the next day, s/he would skip buying the ticket because the chances that inspectors checking the same line two days in a row are perceived as slim.

In the IS treatment, we thus may expect the effect of both sanction and checks on future contribution levels to be ambiguous.

Under collective sanctions, however, the situation is different. Here, if the members of a group get checked, but no one gets punished, this immediately implies that everyone in the group followed the rule. In the CS, being checked without anyone getting punished is thus extremely rich in terms of information. Returning to the train analogy, it would be similar to a situation in which everybody else aboard knew that there is no fare-avoiding stowaway in the entire carriage. In contrast, if the group is checked and punished, this sends a diffuse signal. Someone in the group is free-riding, but it is not clear who, but everyone nevertheless has to bear the consequences. We may therefore expect punishment under CS to discourage future contributions.

In model 3 (Tables 1.3 and 1.4), we included lagged variables of the external check at $t-1$ and external sanctions at $t-1$. These two lagged variables work in opposite directions: if the group is checked, this increases the investment into a group project in the next period, but if it is checked and punished the contributions drop.

Overall, out of 1,620 individual observations, 1,098 (67.78%) were not checked, while 259 (15.99%) were checked without external sanctions, and 263 (15.23%) were checked and punished externally. Therefore, the groups

were checked in 32.22% of the cases, which fits almost perfectly to a predicted 33% level outlined earlier.

Using two binary variables ("External check" and "External punishment"), we constructed a new categorical variable in order to conduct a more fine-tuned analysis. Theoretically, the variable can take 2×2 values. A group can be (1) "not checked, not punished", (2) "checked, not punished", (3) "checked and punished", and (4) "not checked and punished". However, the last option is not realistically feasible option, leaving us with three, rather than four distinct values.

The coefficients of "check, no punishment" and "check, punished" show how deviations from the baseline scenario (no check, no punishment) during the previous period influenced the contributions in the subsequent period. We can see that that subjects reacted differently to external punishment and checks under the two different regimes. Checks of already cooperative subjects (in IS) or groups (in CS) increase cooperation in the next period, even if barely so under IS.

In contrast, the effect of actually being punished goes in different directions for IS and CS. The external punishment of a non-cooperative subject in IS causes his or her cooperation level in the next period to drop. In CS, external sanctions of a non-cooperative group increase that group's cooperation in the next period.

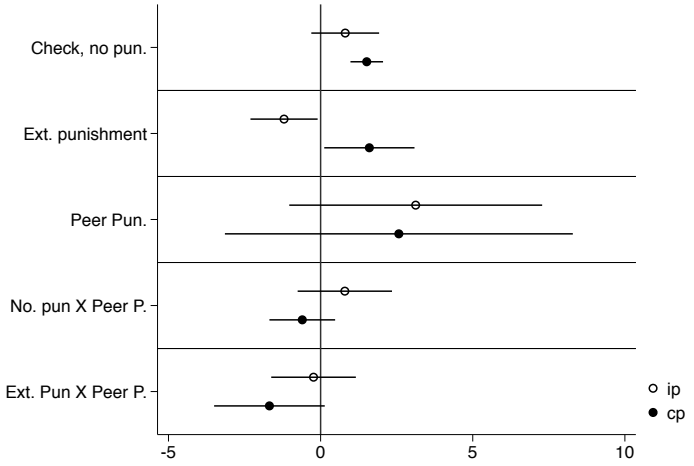


FIGURE 1.11: Panel OLS (Random-effect). DV: Contribution. All IVs are lagged $(t-1)$

We can see that most of the mechanisms discussed above seem to be reflected in the data. In IS, subjects who invested at $t-1$ above the threshold of 11 tokens, by being checked at $t-1$, received a confirmation of their belief that this level of investment is the right strategy. However, we can also find evidence for the "lightning never strikes twice"-logic. Those who decided not to cooperate at $t-1$, and who were fined individually for their non-cooperativeness, supposedly believed that the chances of being checked twice in a row were negligible.

The situation is different under the CS regime. In this case, the previous experience (at $t-1$) of being checked and punished increased compliance. That is, rather than being discouraged by the evidence of the presence of free-riders, and lowering contributions, punishment actually causes contributions to rise. Why does the same "lightning" logic not work here? One possible interpretation is the diffused responsibility under collective sanctions; the group is punished anyway, regardless of the individual decision of a subject, making a strategic violation of the rule harder to calculate. We may also speculate that the increased level of cooperation after the group is punished can be due to Mechanism 1, which we mentioned above: that is,

moral costs imposed on other group members. To draw such a conclusion, we would need more data, however, because the effect of interaction term "Peer sanctions X External sanctions" remains weak.

1.8 Collective sanctions and trust

The declared level of trust affects both the propensity to free-ride, and the level of cooperation. Evidence for a positive correlation between trust and cooperation comes from a study by Gächter, Herrmann, and Thöni, 2004. They revealed this correlation in a large scale lab experiment in Russia, and it has since been replicated many times – for a review of more than 200 studies of this topic see a meta-review by Balliet and Van Lange, 2013.

In the present experiment, participants answered to the standard trust question in a post-experimental survey. We used the formulation from the World Values Survey, which states: "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?", to which there are two possible answers. 40 out of 108 participants (37.04%) chose the answer "Most people can be trusted", while 68 (62.96%) answered that you "Can't be too careful".

In the original formulation the 'Trusting' answer comes first (no randomization of order is done), and coded as 0, and 'No trust' answer is coded as 1, thus in a later analysis negative coefficients in 'Trust' factor should be interpreted as positive effect of trust on a dependent variable.

Whether self-declaratory statements such as the trust question in WVS match with the behavioral dimensions of trust are a question of debate. To adjudicate in this debate, trust measured by survey responses is sometimes compared to the decisions made in a trust game (with a sender and receiver). Some previous studies have shown that the WVS measure correlates with recipient's behavior only (Glaeser et al., 2000), while others have shown the

opposite (Fehr et al., 2003), or no correlation Thöni, Tyran, and Wengström, 2012, using large-scale experimental data ($N = 1500$) in Denmark, showed that in a standard public good game, the average contributions of conditional cooperators (i.e. those participants whose contributions correlate significantly with the average level of contribution in their group) strongly correlate with their answers on the WVS trust question.

The results of our study support the Thoni et al. study. Overall, the contributions of self-declared trusting participants were 32% higher than no-trusting ones (11 tokens vs. 8.3), and this difference is even larger (60% higher or 10.5 vs. 6.5) if we look at treatments without peer sanctions, as in Thoni et al. However, the introduction of collective sanctions eliminated this difference, and the differences in contribution levels become statistically insignificant.

The combination of these findings with gender shows that a lack of social trust is negatively related to contributions among women in all treatments but CP1IS1 It seems that their generally hostile reaction towards CS decreases the cooperativeness even among the most "trustful". No such association was observed among men.

After the experiment, we showed the participants their average contributions over all 15 periods and asked how much, in their opinion, they would have contributed under the alternative regime (that is, CS was shown to those in the IS treatment, and IS to those in the CS treatment). In all cases, participants stated that they would have contributed more in the alternative treatment than they did in their actual treatment (+2.8, S.E. 0.51). The most "regretful" group were those who participated in the CS without peer sanctions. On average, they would have preferred to invest 5.2 tokens more in the group project under the alternative regime, S.E. 0.98).

1.9 Peer punishment across treatments

This section analyzes how the use of peer punishment varied between the different sanction regimes and different genders. We should note that due to a relatively high cost of punishment (1 token per 2 deduction points), the number of instances of peer punishment was quite low: out of 51 participants, only 33 ever used it, and did so in only 84 cases. This limits our ability to draw statistical inferences by "traditional" methods, such as a panel OLS.

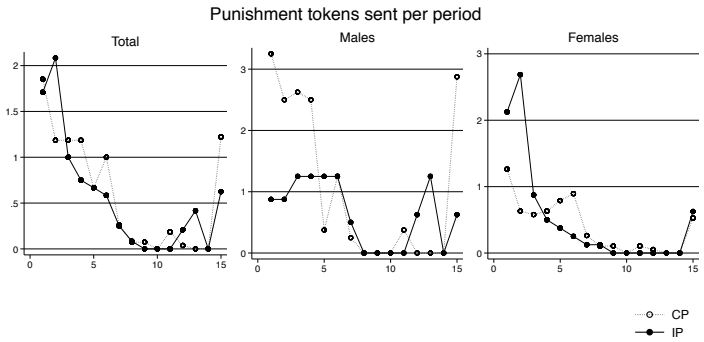


FIGURE 1.12: Peer punishment: general and across gender

Nevertheless, the patterns observed are intriguing. Initially, the dynamics of peer punishment are almost perfectly the same between CS and IS. In both IS and CS treatments with peer sanctions, the peer punishment dies out after the 7th round, having fulfilled its cooperation-stabilizing role. If we look closely at the punishing behavior of males and females separately, we see two rather different stories emerge, however. Males punish their group members much more under collective sanctions, while females do so under individual sanctions. We thus here have another piece of evidence that the different genders react systematically different to the different sanction regimes.

1.10 Conclusions and implications

Despite the positive evaluation by some policy makers, collective sanctions do not seem to result in higher levels of group cooperation. The mechanism of delegation does not work as it should. Its failure can be explained by strong negative feelings against the entire regime of collective sanctions – the sentiment that an approach that treats the group as one entity is deeply immoral.

Another reason why collective sanctions fail may be found in a wrong incentive structure. In dynamics across all 15 rounds the cooperation rate under collective sanctions have been lower than under individual sanctions. One of the plausible explanations is the main mechanism that presumably had to drive peers to punish their team members who do not cooperate, works in the opposite direction. Collective sanctioning sends a mixed signal punishing cooperators despite their pro-social behavior. Instead of looking for a cause of their trouble within the group, cooperators start changing their behavior towards defection.

This feeling is mostly shared by those who position themselves as trustful. Thus, lacking an institutional justification, collective sanctions might erode social trust, and, as a result, decrease cooperation instead of increasing it.

To make collective sanctions work, a greater effort may be needed to make people believe that their use is fair. In order to avoid framing effects in the lab, we used highly neutral wording to describe the collective sanction regime. We thus applied the, arguably unrealistic, assumption that participants would need no information on the origin and "justifiedness" of the specific regime. In real-life situations, the introduction of such a fundamentally unfair regime begs for a certain level of justification and legitimization. When used in practice, collective sanction regimes are usually explained by the difficulty of detecting individual wrongdoers, and the wish to delegate

such detection towards their peers. That is the way collective sanctions usually work in school collectives; for example, if no one admits to breaking a window, the entire class is kept after school. Here, the central authority tries to change the structure of the costs and benefits of whistle-blowing.

Another approach to justifying collective sanctions is with reference to the guilt of third parties – a guilt that precedes the "wrong" action by an actual, individual infringer, and which may be sees as having caused the later infringement. The Russian tsarist government explained their use of collective punishment to peasant communities for tax evasion of individual members in this way. All peasants of the community are guilty, the logic goes, for "not trying to convince their fellow to fulfill his duty, when they observed him giving himself up to negligence and idleness" (Brockhaus and Efron 1896 Article "Frankpledge"). Different legitimation discourses may thus have an effect on the efficiency of collective sanctions. Additional studies on the subject will be necessary to study these.

The generally higher level of cooperation shown by male subjects may be explained by the higher cultural acceptance of collective sanction regimes in predominantly male collectives, such as prisons and the army. This hypothesis should be checked either in the field or in the lab. If we can induce the permissibility of collective sanctions in the lab, would it change the attitude towards it, and promote higher levels of cooperation?

As for the gender effects, we do not yet know if these occurs only in gender-heterogeneous groups (the case of the present experiment), or whether different effects would be observed in gender-homogeneous groups. An experiment should be performed to determine how male subjects react towards the application of collective sanctions on females – and whether they would continue to support this sanction regime in the same way they do when it is applied on themselves. It is probable that the idea of applying collective sanctions on women is culturally unacceptable for male participants as well,

and that it would provoke a counter-reaction, similar to what we have seen with female participants.

Limitations of the present experimental design for external validity notwithstanding, we believe that the results presented here make an important contribution to the existing body of knowledge about collective sanctions. Their introduction does not automatically increase public good contributions. Neither do collective sanctions make horizontal sanctions more efficient. Taking into account that collective sanctions can be more costly, produce a deep sense of unfairness, and can lead to the rebellion (Heckathorn, 1990), they should not be considered an effective measure against free-riding or other kinds of antisocial behavior, at least not without any significant legitimation.

Chapter 2

Can intergroup collective sanctions increase cooperation between groups?

Abstract

When members of one group encounter a norm violation committed by a member of another group, this antisocial behavior is often handled by picking a random member of the community to which the perpetrator belongs and by applying sanctions to him/her. Despite its prevalence, this kind of third-party sanctioning is puzzling. Sanctioning for norm violations is a public good, so there is always a temptation to free-ride, hoping that someone else will enforce the norm. From a moral perspective, too, collective sanctions are puzzling as they are not actually aimed at a guilty person. Nevertheless, in real intergroup conflicts, collective sanctions are widely used. Functionalists suggest that groups resort to collective sanctions because this increases the degree of cooperation between them. This effect is achieved by the delegation of (peer) punishment from outgroup to ingroup members. This study tests whether collective sanctions applied by outgroup members result in higher intergroup cooperation, and whether the introduction of collective sanctions increases the amount of ingroup punishment. The results demonstrate that neither of these two functionalist arguments come true: participants avoid using collective sanctions against outgroups, and the amount of ingroup (third-party) punishment is no higher under the intergroup collective sanctions regime.

Keywords: Collective sanctions, Public good game, Crime deterrence , Intergroup conflict

2.1 Introduction

Everyday life provides a plethora of examples of when members of one group are oppressed or harmed by another group indiscriminately as a response to a crime or norm violation committed by an individual. Despite this, the prevalence of intergroup conflict is significantly lower than is predicted by most models. The functionalist argument states that the institution of collective sanctioning serves as a tool that promotes cooperation. There are two mechanisms that make such sanctioning effective: first, a person may be driven to avoid bringing harm upon their own group members because of some sort of kinship altruism, or, second, random punishment increases the chance for ingroup policing. The aim of this paper is to check experimentally what happens to intergroup cooperation when a person can retaliate against a random member of another group as a reaction to an individual's norm violation.

To begin, we will look into two situations in which indiscriminate intergroup sanctioning took place; these will serve as a starting point for further considerations of intergroup collective sanctions' efficiency.

In August 26, 2016 in Loschinovka, a small town in southern Ukraine, a 9-year-old girl was murdered. The local residents suspected Mikhail, a 21-year-old friend of the family and member of the Roma community to be the perpetrator. Two days later, at a community gathering, the citizens of Loschinovka made the decision to expel the entire Roma community from their town. Despite attempts by the police to prevent this, the eviction was executed the next day. According to the local press, a year later, in September 2017, the expelled families still remained homeless, mostly sheltered by their relatives in nearby towns. This tragic story exemplifies the use of intergroup collective punishment: as a reaction to a crime committed by a single member, the entire group was punished.

Often, intergroup collective punishment takes a slightly different, more individualized form, however. A few days after the terrorist attacks of 9/11, a Boeing aircraft mechanic from Arizona, Frank Silva Roque, drove with his truck to the petrol station belonging to Balbir Singh Sodhi. Roque shot Sodhi five times with his handgun, killing him. Roque was driven by the itch for revenge on all Muslims for what had been done by Al-Qaeda in the September 11 attacks. As he admitted to his friends that day, he was "going to go out and shoot some towel-heads." The sad irony that his victim's religious affiliation was Sikh should not hide the curious fact that Roque decided to kill a random Muslim in retaliation of the 9/11 attacks.

These two cases are structurally different. The first case was a collective decision of one group (the villagers) to impose sanctions on all members of another group to which the suspect belonged. Both groups were relatively small, and the sanctioning group could blame the local Roma population for inaction and indirect commission of this crime. Both groups had cohabited together for a long time and could thus assess how close ingroup members were. In contrast, the murder committed by Roque was driven by an individual desire, and was aimed at a random person who unluckily was perceived to have the necessary feature of being a Muslim. What unites both cases is a desire and demand for what psychologists call displaced revenge: third-party punishment directed to an entire group or a random representative of that group. In this paper, I will make one step towards a deeper understanding as to how displaced revenge can affect intergroup relations.

We may brush off both cases simply as examples of "impulsiveness, irritability, incapacity to reason" of the crowd (Le Bon, 1897). However, similar sanctioning mechanisms are also suggested by serious academics. In a recent study by Ginsburg and Simpser, 2017, the authors suggest using collective sanctioning as a way of controlling recent migrants' behavior in the US. They suggest that migrants wishing to enter the US should form "trust circles",

which will be collectively held responsible should one member of such a circle commit a crime or other infringement. According to the authors, such a policy would have two main consequences: a positive selection effect, and increased ingroup policing. People who take the decision to migrate via the "trust circle" policy will be cautious in selecting their partners in a way that their partner's delinquent behavior will not undermine their perspective to stay. What is more, after migration, members of "trust circles" will be more inclined to detect and inform on group members when they observe them engaging in suspicious behavior.

The unfairness of such a way of rendering justice is obvious: the entire foundation of modern justice is based on the concept of individual responsibility. No one would like to be held responsible for something s/he did not do. Nevertheless, as shown, intergroup collective or random sanctioning is common, and, for some, has strong intellectual appeal. This paper looks at the consequences of intergroup sanctions by asking: how do intergroup collective sanctions influence intergroup relations?

2.2 Perspectives on intergroup cooperation

Both sociologists and social psychologists have quite low expectations regarding the prospect of intergroup cooperation. In general, they project the Hobbesian war logic to the group level: groups should permanently be competing over limited resources and should thus be in permanent conflict with their neighbors. The classic 1960s Robbers' cave experiment seems to support this intuition: the two groups were in strong competition throughout, and were only able to fulfill a joint cooperative task when they had a goal that united them both into a larger "supergroup" (Sherif, 1961). In fact, from the sociological perspective, intergroup conflict may be constitutive for a group,

as violence towards strangers can provide a sense of unity to a group (Gould, 1999).

Many political scientists hold the assumption that without a state that controls and intervenes, heterogeneous groups are doomed to be driven into conflict, and there is no chance for peaceful intergroup cooperation (see for instance chapter 14 of Horowitz 1985). Evolutionary game theorists show that altruism (putting a value on the utility of another person even if this is costly to yourself), which we value so highly in interpersonal relations, is closely tied to parochialism (hostility towards those who do not belong to your group), and evolutionarily speaking, these two traits developed simultaneously (Choi and Bowles, 2007).

There is another, more optimistic line of thought, though. Its adherents point out that the empirical facts contradict the predictions of permanent intergroup group conflict, both for 'primitive' societies and for modern ones. Many anthropologists and sociologists demonstrate that in many stateless societies, intergroup cooperation, such as trade, flourishes. The obstacle of mutual distrust is often solved through the development of complicated systems of gift exchange (Landa, 1994). In his historical overview of intergroup cooperation in the Middle Ages, Leeson, 2008 also rejects the idea that without the state, groups are doomed to intergroup conflict. For instance, the analysis of court documents in Corsica, where presumably the tradition of feud was widespread until the 19th century, showed that feuds were in fact rare. In the modern world, if we compare the *actual* frequency of interethnic conflicts in Africa with the number of *potential* interethnic conflicts, this ratio does not even reach 0.1% (Fearon and Laitin, 1996).

2.3 Why is intergroup cooperation problematic?

Before considering how collective or random sanctioning may convince members of different groups to cooperate, let's briefly review the main reasons why cooperation fails. There are several reasons why intergroup cooperation can be problematic. First, there is information asymmetry between groups. The reputation of outsiders is usually unknown, and the costs of obtaining this kind of information are higher compared to those for getting the same information about members of one's own circle. Second, ingroup bias undermines the expected reciprocity in cooperation. When an individual meets a member of another group, they both know that the chances of cooperation are lower than with their own group members, and that, in turn, the person has a lower expectation of reciprocity. Finally, a person who, due to a bad reputation, fails to cooperate with their own kin has to look for outside contacts. The very intention of someone to deal with an outside group can be a sign of their untrustworthiness, known inside the group. This negative selection effect is closely related to information asymmetry.

How displaced revenge may solve the intergroup cooperation problem

If, despite all obstacles generated by mutual distrust, lack of information about the agent's reputation, ingroup bias, and negative selection, we still observe intergroup cooperation more often than war, there should be a mechanism that curbs intergroup conflict and allows people to cooperate across group borders without fear of being betrayed by the partner. As functionalists suggest, the mechanism in action here is the entitativity, or non-distinction, between the individual members of the outside group. As we could see in this paper's introductory examples, people often treat members of an outside group as a whole, without distinguishing between individuals. A mob killing an innocent immigrant because another immigrant committed a crime one day before is among the most deplorable examples of such

stereotyping, which nevertheless occur quite regularly.

There are two explanatory layers that account for the existence of random sanctioning: individual-based and group-based, or, in other words, psychological and functionalist explanations.

2.3.1 Psychological explanation of collective sanctions

From a psychological perspective, random or collective intergroup sanctioning falls within vicarious or third-party retribution. It is a situation in which "a member of a group commits an act of aggression toward the members of an outgroup for an assault or provocation that had no personal consequences for him or her, but which did harm a fellow ingroup member. Furthermore, retribution is often directed at outgroup members who, themselves, were not the direct causal agents in the original attack against the person's ingroup." (Lickel et al., 2006).This definition is insightful for this paper because it introduces both the third-party punishment, where the punisher is not the target of a norm violation (Fehr and Fischbacher, 2004), and collective sanctioning where the punished individual is not the source of such a violation.

There are two lines of thinking that seek to explain this phenomenon at the level of the individual. The first one links intergroup conflict with outgroup entitativity (Newheiser and Dovidio, 2015), i.e. the perception of a group as a pure entity abstracted from the individuals it is composed of. The larger the extent to which the group is perceived as unified, the higher is the intention to punish the whole group or one of its random representatives (Newheiser, Sawaoka, and Dovidio, 2012). When outgroup entitativity is high enough, displaced punishment can be as satisfying from a justice point of view as direct punishment (Sjöström and Gollwitzer, 2015).

That provides a rational explanation for such acts as the murder of the innocent Sikh Balbir Singh Sodhi by Frank Roque. A punisher cannot (or does

not want to) distinguish between specific group members, and treats individual defections as a representation of the behavior of the entire group. A rational agent increases the punishment contributions when the cost of punishment drops (Nikiforakis and Normann, 2008), so it is logical to choose a target for sanctioning that is associated with the highest punishment efficiency, for instance the first available member of the outgroup. Since it would be much costlier for Frank Roque to find and punish a real culprit of the 9/11 attack rather than attacking Balbir Singh Sodhi, who was the first in his way, he chose to punish the latter.

Instead of hypothesizing about the non-distinguishability of different outgroup members, another approach would be to build a causal chain projecting the individual guilt to those who are not personally involved, but just share their group membership with the perpetrator. This is the so called a "gatekeeper" logic (Kraakman, 1986), which is behind all types of third-party or vicarious responsibility in modern law. Group members are punished not for the action that was committed by their group member, but for their inaptitude to prevent this action. Formally, 'gatekeeper liability' is defined as "liability imposed on private parties who are able to disrupt misconduct by withholding their cooperation from wrongdoers" (Ibid. p. 53). So, in simple terms, group members are not as innocent as they claim, because they could prevent a crime toward another group if they wanted to. As psychological vignette experiments have shown, this legal line of thought is used by the general audience as an argument for displaced revenge. People tend to blame outgroup members either for indirect commission – the presumption that other outgroup members encouraged the antisocial act, or omission – the failure to prevent the act (Lickel, Schmader, and Hamilton, 2003).

2.3.2 Functionalist explanations of intergroup cooperation

What if collective punishment can provide stability and peaceful co-existence for two groups living side by side? Fearon and Laitin, in their 1996 paper on interethnic cooperation, made exactly this argument. The logic is simple: the fear that a person's individual misbehavior towards an outgroup may provoke full-scale intergroup conflict averts people from harmful actions.

One of the possible strategies that makes intergroup cooperation sustainable in the long run involves collective responsibility of the entire group if one of its members misbehaves: "...to support cross-group cooperation solely by threat of punishment by the offended group, punishment must be indiscriminate, targeting either all members of the cheater's group or some random collection of them" (Fearon and Laitin, 1996, p.722). Collective sanctioning moves the scope of the conflict from the individual level to the collective by involving other parties not part of the initial contract. The threat of such intergroup conflict – effectively a threat of cycles of retaliation between the groups – will make participants more reluctant to defect with members of the outgroup. In other words, collective sanctions make short-term profits reaped by defection less attractive because the intergroup conflict the defection produces will be hazardous.

Moreover, collective sanctions put innocent ingroup members under the threat of intergroup sanctions, which, in turn, will increase their motivation to engage in ingroup peer punishment. This argument certainly applies to kin. Many parents would prefer to suffer themselves rather than have their children punished. The fact that altruism towards kin can increase levels of cooperation is demonstrated by the institution of hostage-taking. Hostage-taking was widely practiced in the Middle Ages to guarantee the fulfillment of contracts, but is also used in North Korea, where diplomat's families are temporarily arrested while the diplomat is visiting other countries in order

to guarantee that s/he will not run away. This kind of altruism, however, is unlikely to apply to larger groups, and it's effect thus limited.

Another rationale behind collective sanctions stems not from an *ex ante* perspective (as the third-party liability argument), but flows from an *ex post* perspective: if a perpetrator from another group is hard to detect and punish, then the task of such punishment should be delegated to those whose opportunity costs of such punishment are lower, which are his own kin or team members. Collective sanctions in this case, despite their unfairness, serve as a delegation mechanism (Hechter, 1988). This delegation solves simultaneously two issues of intergroup cooperation: ingroup bias, and information asymmetry.

2.4 Ingroup bias and third-party punishment

Psychological experiments on discrimination (Tajfel, 1970) showed that even the weakest identification with a group results in discrimination towards outsiders. There is a long debate about what is the source of this discrimination. While the traditional view was identity-based, more recent behavioral experiments demonstrated that the root of discrimination lies in general reciprocity: group members favor their own co-members because they expect that they will return the favor in the future (Yamagishi and Kiyonari, 2000). This idea can be modelled in terms of indirect reciprocity (Nowak and Sigmund, 2005). The indirect reciprocity model suggests that people cooperate with each other because of reputation concerns. The reputational feedback loop can be split into two parts: upstream and downstream reciprocity. In upstream reciprocity, a person who recently received some benefits from a third party feels that s/he would like to do the same to others. Later, in downstream reciprocity, the benefit is paid back when a person who knows that this individual has a reputation to treat others well will return him this

favor. This logic of indirect reciprocity is based on the assumption that there is a public space where information about an individual's previous interactions is freely shared. By definition, this sharing of reputation is hindered across group boundaries. We treat outside group members as one unit; thus, one member's bad reputation casts a shadow on the reputation of everybody else in the group. Information exchange is less intense across groups, so the transaction costs to inquire about this specific person are higher when the person is from an outgroup.

We may expect that ingroup bias would push people to punish their own group members less than outsiders when they violate a norm. In reality, the situation is more complicated. Some studies confirm the ingroup leniency hypothesis (Lieberman and Linke, 2007): that people tend to be more tolerant to their own group members. However, if we treat punishment as a second-order public good, we should expect the black sheep effect: people tend to produce more of it when they deal with the ingroup members (Shinada, Yamagishi, and Ohmura, 2004). Indeed, some studies find the "black sheep" effect in third-party punishment; the willingness to punish offenders is increased if people deal with their own group members (Gollwitzer and Keller, 2010). Finally, some field studies demonstrate even more complex patterns.Experimental research in Papua New Guinea showed that norm violators expected that a norm enforcer belonging to their group will be lenient to them. At the same time, this research showed that punishers cared about was not the group membership of perpetrator, but only that of the victim of the crime (Bernhard, Fischbacher, and Fehr, 2006).

Whatever effect (black sheep or ingroup leniency) we observe in reality, the introduction of intergroup collective sanctions presumably should increase the frequency of ingroup punishment. For instance, in an intergroup trust game, two different institutional mechanisms were tested: ingroup punishment and sharing the information about defectors to outsiders

(Kimbrough and Rubin, 2015). The authors found a marginal effect of information sharing, but the introduction of ingroup punishment substantially increased intergroup reciprocity.

Greif, 2004 provides a historical perspective of a sanctioning system with a significant ingroup bias. In Medieval Europe, the judicial system always took the side of the local citizen in his or her argument with an outsider. In order to deal with such a bias, the community responsibility system (CRS) was developed. This system makes all members of the outside community responsible for their own members' violations of a contract. The norm was enforced through the seizure of property from any outside-community member who happened to be in the jurisdiction of the community of the victim. Thus, there were two ways to react for the outside community: either to stop any kind of activity with the local community, or to detect and extradite the norm violator. As described by Greif (2004, 2006), over time the second solution gained in traction in the form of merchant guilds and community responsibility systems (CRS). In guilds, traders and craftsmen enforced norms of conduct within their communities, and also punished misdeeds carried out in other cities by their peers. In return, guilds or CRS in that city would do the same with their own members. This system of mutually interdependent contract fulfillment functioned without any involvement of the central state, which was very weak and prejudiced against outsiders.

2.5 Information asymmetry and third-party punishment

A key issue in intergroup cooperation is information asymmetry. Within a group, the cost of obtaining information about an individual is low. Punishment strategies can be based on easily-observable individual behavior. However, when actors from another group decide to interact with members of the first group, they often find themselves in a disadvantageous position: it is hard to enforce a contract and even to identify the counterpart if the person breaking it belongs to another group. Being unable to "get even" with specific individuals, the actors have to deal with the entire group, which can result in discrimination. This is one of the main reasons why integroup cooperation is so often based on stereotypes: "if individuals are hard to identify or investigate across groups, then intergroup cooperation and trust cannot be supported by punishment strategies that condition on individual behavior" (Fearon and Laitin, 1996). If ingroup members know significantly more about the perpetrator than outsiders, that knowledge solves the problem of asymmetry. Outgroup members can now delegate the responsibility to find and punish the norm violator. A final question is who among the outgroup members should be made responsible to enforce the norm: "Groups can delegate the monitoring task to individual agents, but this is likely to entail high costs. If the burden of monitoring could be shared among the entire membership, then agency costs could be avoided. The rub is that rational members will only engage in monitoring on the group's behalf if they have a sufficiently large incentive to do so." (Hechter, 1988). Here, intergroup collective sanctions serve as a "large incentive" that can help to solve the internal collective action problem.

2.6 When intergroup collective sanctions fail

The main problem with collective or random sanctioning is the mixed signal that it sends to the lawful members of the outside community. This problem is made clear by studies on antisocial punishment – punishment meted out against cooperators rather than free-riders. Numerous studies across the world have demonstrated that antisocial punishment is common (Herrmann, Thöni, and Gächter, 2008), and that when cooperators are punished, this undermines cooperation (Rand and Nowak, 2011). Collective or random sanctioning can be likened to antisocial punishment since by definition, such sanctions target individuals that are not to blame. Thus, in a similar fashion to antisocial punishment, we may expect that the more often cooperators are punished for the actions of their co-members, the less incentive they have to cooperate in the future. This makes collective sanctions sensitive both to group size and the frequency of interactions. As shown theoretically (Heckathorn, 1988), collective sanctions are efficient only in very small groups. Historical evidence confirms the theory; merchant guilds and community responsibility systems worked quite efficiently for propagating trade and cooperation between medieval European communities. Ironically, their efficiency was the reason for their collapse. As international trade and the frequency of intergroup contacts grew, so did the incentive to fake community membership to gain from its reputation. Without modern means of communication, the chances were high of punishing communities that had nothing to do with perpetrators (Greif, 2002).

Collective sanctioning face issues on the punisher's side as well: people tend to avoid punishing a yet-to-be-identified wrongdoer (Small and Loewenstein, 2005). That reluctance is caused by the same logic as the "identifiability of the victim" effect. We feel more sympathy to a specific person than to an abstract one, and we dislike the specific wrongdoer more than an

abstract one. Under random sanctioning, the target is unknown, so the willingness to punish is lower. Collective or random sanctioning also violates one of the main postulates of any punishment: it should be fair. The demand for fairness has been found to be the leading factor defining the willingness to punish (Singer and Steinbeis, 2009). This is another reason why collective sanctioning is sensitive to group size: under collective sanctions, the chances that punishment will hit the specific participant whose misdeed has been observed declines with the size of the group. Ingroup members will therefore be less willing to collectively sanction an outgroup the larger in size that group is. The declining willingness to punish can provoke a chain reaction, resulting in lower cooperation and higher rates of defections:

1. The expectation of lower chances of being punished increases the propensity to defect in the outgroup.

2. The lower level of expected cooperation among outgroup members decreases the chances of being held accountable by the outgroup, thus undermining the willingness to punish ingroup members as well.

3. As a consequence, the lower chances of being punished for defection by both in- and outgroup members decreases the level of cooperation within and between groups.

From the above it should now be clear that, theoretically, intergroup collective sanctions may be associated with both *increased* or *decreased* levels of intra- and inter-group cooperation. The paper now moves on the empirical section. As it unfortunately is impossible to test all theoretical conjectures mentioned above in a single experiment, I restrict myself to testing a core prediction that emerged from the discussion. Specifically, I test the hypothesis that collective sanctions increase the frequency of in-group third-party punishment.

2.7 Experiment

To test empirically the effectiveness of collective sanctions in enforcing cooperation, I designed a dedicated lab experiment in which I compare participants' behavior under two different sanctioning regimes. The first regime combines intergroup collective sanctions with peer punishment and is henceforth referred to as intergroup collective sanctions (ICS). Under this regime, participants can punish out-group members collectively, but also have the option to exert influence over their peers by means of peer punishment. I benchmark collective sanctions against another sanction regime: integroup individual sanctions (IIS). The individual sanctions regime allows *individual* sanctioning (i.e. peer-punishment) not only of in-group members, but *also of out-group members*. This individual sanctions regime therefore gets rid of the fundamental problem of intergroup norm enforcement, namely the information asymmetry between ingroup and outgroup members.

I use these two treatment conditions to test how effective intergroup collective sanctions are in terms of encouraging cooperation relative to the 'gold-standard', intergroup individual sanctions. Specifically, I inquire how overall rates of cooperation and punishment behavior differs between the two conditions. Specifically, I hypothesize that 1) cooperation rates will be higher under IIS, and that 2) the lower cooperation rates under ICS can be attributed to a reluctance to (collectively) punish members of the other group under ICS. Put another way, we should see relatively higher levels of (individual) intergroup sanctioning in IIS, which can explain the higher cooperation rates under IIS as compared to ICS.

2.7.1 Experimental design

Two of the main theoretical works analyzing intergroup cooperation, Fearon and Laitin, 1996 and Stoff, 2006, both use a social matching game (Kandori, 1992) as their baseline model. In this game, members of two different groups are matched with each other, and in each round, they play a Prisoner's Dilemma game. As Fearon and Laitin showed, one of the subgame perfect Nash equilibriums is to cooperate until one member of the outside group defects. Then the entire group should stop cooperating completely with outsiders. Stoff proved theoretically that this strategy produces a stable equilibrium even faster when combined with ingroup punishment.

In this paper's experimental design, I made a slight modification to Stoff's game. My goal here is to test cooperation rates in two different sanctioning settings. In both cases, people first play Kandori's social matching game with a partner of the other group. Then they observe the decisions made by another matched pair and make decisions on third party punishment. In two different treatments two various institutions of sanctions were in action. In the Individual intergroup sanctions (IIS) condition, participants can punish the specific ingroup member involved in the interaction they observe, *and* the specific outgroup member while under collective sanctions a **random** outgroup member is punished.

The experiment was conducted online with users of Amazon's Mechanical Turk platform. Data was collected in two separate sessions on September 13 and 14, 2017. The first session was run as the 'Individual intergroup sanctions' treatment, and the second session as the 'Intergroup collective sanctions' treatment. In each session, 60 participants took part.

As explained shortly, participants took part in the experiment in groups of 6. Since workers on Mechanical Turk select their assignments individually,

it was not possible to recruit the required groups of 6 in one go. This problem was solved in the following way: After signing a consent form, MTurk workers were redirected to a waiting room, where they waited for 5 other members to join. In order to avoid drop-outs, the participants were compensated for waiting at a rate of 10 cents per minute of waiting. After those 10 minutes, participants would receive their participant fee, their compensation for waiting and were free to leave the study. In this way, the problem of initial high-dropout rates was entirely solved. Unfortunately, one participant dropped out in the middle of game, which meant that an entire group of 6 was not able to finish the task. The total number of participants whose decisions are used in the analysis therefore is 114.

Once 6 participants had joined the waiting room, they were assigned to a super group based on their arrival time. Each such group was subdivided into two subgroups of equal size of 3, named Group A and Group B. Participants were informed of their group membership and that they would keep this membership across all 20 rounds.

Despite the fact that group membership remains fixed across rounds (a partner matching), the identities of specific members were left unknown in order to avoid strategic punishment, thus group members were not identifiable to each other between rounds.

The game consisted of two stages.[1] In the first stage, participants engaged in a Prisoner's Dilemma game with members of the other subgroup. Each participant from Group A was randomly matched with a participant from Group B. So in each round, three pairs of players were created.[2]

[1]Instructions for both stages were shown after assignment to the groups, and remained available throughout the entire study. Screenshots of the instructions are included in Appendix F.)

[2]Since the subjects stayed part of the same super group of 6 for the entire set of 20 rounds, the game can be considered a repeated Prisoner's Dilemma, despite the fact there was only a 1/3 chance in each subsequent round to be matched with exactly the same player (cp. Hennig-Schmidt and Leopold-Wildburger, 2014).

Each participant now had to decide how much out of his/her endowment of 10 points to send to their partner from the other group. That is, participants had the option to send anywhere between 0 and the full endowment to their partner, knowing that their partner would take the same decision simultaneously. I thus followed the lines of Capraro, Jordan, and Rand, 2014, who played a set of continuous prisoner's dilemmas with various benefit-to-cost ratios of cooperation. In comparison to the dichotomous outcomes in traditional binary Prisoner's Dilemma games, this continuous scale allowed me to get the more nuanced measures of cooperation. Whatever amount participants sent to their partner was multiplied by a factor of 2.

Game-theoretical predictions

The predictions for the **subgame** that consists of a Stage 1 only are the same as for the traditional finite Prisoner's Dilemma with dichotomous choices: a short-term profit maximizer would send nothing to their partner, while at the same time hoping that their partner would cooperate fully. Since both partners would apply this same logic, however, the theoretical prediction is full defection. This is despite the fact that both partners could double their payoffs if they reciprocated and sent each other their full endowment.

The theoretical predictions for the entire game (Stages 1 and 2) are delineated in Fearon and Laitin (1996). The entire system has a chance to move from full defection when the opportunities for either collective sanctions or in-group punishment are introduced. There are three possible outcomes. First, the full defection is still an option as in the subgame with no sanctioning institutions. Since punishment and external sanctions incur additional costs for a punisher, that creates a motivation to avoid these costs. Another equilibrium is a spiral one: out-group members punish randomly members of another group, as soon as anyone encounters the defection in Stage 1 (Prisoner's Dilemma), and this punishment triggers a counter-punishment wave

from another group. However this stage in a *spiraling equilibrium* is never reached because each party in Prisoner's Dilemma avoids defection expecting higher losses in case of an intergroup conflict. The last possible theoretical outcome of a full game is what Fearon and Laitin call *in-group policing*: no defection in mixed (outgroup-ingroup) interactions is punished, but the defectors are punished internally by their own group members. In both latter cases the equilibria result in full cooperation in Stage 1. It is crucial to emphasize here that both equilibria are purely 'theoretical' in a sense that in a real life a combination of these two polar solutions would be observed: a threat of outward collective sanctions result in the in-group policing to some extent.

In the second stage, participants were given the option to engage in third-party punishment of both ingroup and outgroup members. This punishment stage was implemented as follows. Immediately after having taken the decision how much to send to their partner (Stage 1), but still *before* they had learned the outcome of their own personal interaction, they were shown the outcome of *another* interaction between an ingroup and an outgroup member that took place in that round. Participants could then decide whether to punish one, both or none of these other two participants. They thus engaged in a third-party punishment decision in the sense of Fehr and Fischbacher (2004) (although these authors studied punishment in the context of a Dictator game, not a Prisoner's Dilemma). Only after participants had decided on the whether and how much third-party punishment to assign were they informed about the outcome of their own personal interaction in the Prisoner's Dilemma.

I decided to have participants engage in third-party punishment – rather than simply have them punish the person with whom they had interacted in Stage 1 – in order to be able to interpret their decision as *norm enforcement*. There are three arguments to be made here. First, if a participant could only

observe and punish his or her personal interaction, he or she may use the punishment stage merely as an opportunity to change the outcome of the previous interaction. Namely, s/he may try to equalize payoffs. As previous studies on direct punishment in public good games have shown, inequality aversion is one of the dominant motives in altruistic punishment (Masclet and Villeval, 2008), so such behavior is quite likely. Second, in dyadic game there is no space for second-order free-riding, making punishment less informative. If you are the only person who interacts with the defector, no one but you can punish him. Third, in direct punishment, the motive of enforcing a norm can be mixed up with the motive of personal retaliation. In contrast, if a person observes the outcome of a third-party interaction, s/he sends a signal about what kind of normative behavior s/he finds appropriate in a group.

Punishment was conducted by means of deduction points that participants could assign to each other. The punishment factor was three, which is the standard setting for public good games with a punishment stage (Nikiforakis and Normann, 2008; Fehr and Gächter, 1999). This means that for each deduction point sent, the recipient's payoff was decreased by three. In the beginning of each the punishment stage, participants received an extra endowment of 10 points. This was done so that the players' decisions on punishment would not be affected by the amount of available points left from Stage 1.

The two treatment conditions of the experiment – the 'Individual intergroup sanctions (IIS)' treatment, and the 'Intergroup collective sanctions (ICS)' treatment – varied with regard to who could be targeted during third-party punishment of outgroup members. In the IIS treatment, the punishment was applied to the outgroup member whose decision was shown to a participant, whereas in the ICS treatment, a random member of the outgroup was punished. Otherwise the two treatments were identical. Our main interest is

how third-party punishment towards *ingroup* members varies between the two treatment conditions. That is, our focus is on ingroup policing in the sense of Fearon and Laitin, 1996 and Stoff, 2006. Beside this, another quantity of interest are the overall cooperation rates achieved under the different punishment regimes.

Before presenting the results, let me discuss once more the possible reasons that could drive participants to send punishment points to their ingroup members. After all, punishing ingroup members may look counterintuitive because participants are never actually matched with them during the Prisoner's Dilemma. By design, they cooperate or defect against the members of the other group only. But as Fearon and Laitin predicted in their model, concerns about group reputation should be a powerful driving factor behind punishment intentions towards your own group. If a participant observes their own group member defecting in the Prisoner's Dilemma, s/he realizes that this defection will lead members of the other group to form expectations of low reciprocity for the entire group for the future PD interactions. Thus, each defector decreases the future chances for cooperation for each member of a group. What is more, defections by another ingroup member may trigger retaliation by the other group, which may negatively affect the participant in focus, even though s/he cooperated.

2.8 Results

Above, it was hypothesized that 1) cooperation rates would be higher under IIS, and that 2) the lower cooperation rates under ICS can be attributed to lower rates of *intergroup* punishment in ICS as compared to IIS. Table 2.1 sums up the overall results. I discuss the various findings in turn. I start with overall cooperation rates, and then discuss punishment behavior.

TABLE 2.1: Results overview intergroup punishment experiment

Treatment	N	Average Contribution	Average ingroup punishment received	Average intergroup punishment
ICS	60	4.94	1.8	1.2
IIS	54	6.20	1.7	2.3

As hypothesized, the overall level of cooperation was lower under intergroup collective sanctions, with an average contribution in all 20 rounds of 4.94 points. This compares to average contributions under individual intergroup sanctions 6.20. This difference is highly significant in a t-test (t-value=6.38, p<0.001%).

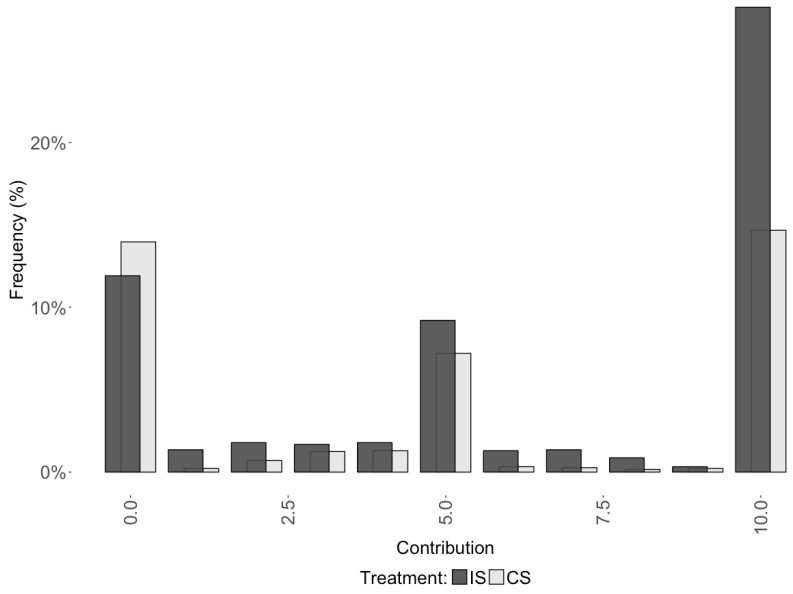


FIGURE 2.1: PD Decisions under IIC and ICS

What is driving this difference becomes clear when we look at the overlapping frequency histograms of PD decisions (Figure 2.1). Here we see that the number of free-riders (those who contribute nothing) is almost the same across treatments. The number of half-contributors (contributing about half of their endowments) is also similar. But the number of those who fully contribute 10 tokens out of 10 is twice as high under the individual sanctions treatment.

As shown in Figure 2.2, both punishment regimes were effective in stabilizing cooperation over time. Contributions started somewhat above the mean of endowment (at 6.9 tokens in the case of IIS, and 6.4 tokens in the case of ICS). That is, remarkably, we do not see the deterioration of contributions over time that is commonly observed in similar games without punishment (cp. Ledyard 1995). In the last round, as it is often the case, we observe an end-game effect with a decline in cooperation rates (cp. Normann and Wallace, 2012).

We can also see that in all 20 rounds, IIS contributions exceed ICS contributions. Second, there are no fully contributing groups in ICS, and there are no groups in IS where contribution level entirely collapsed to zero.

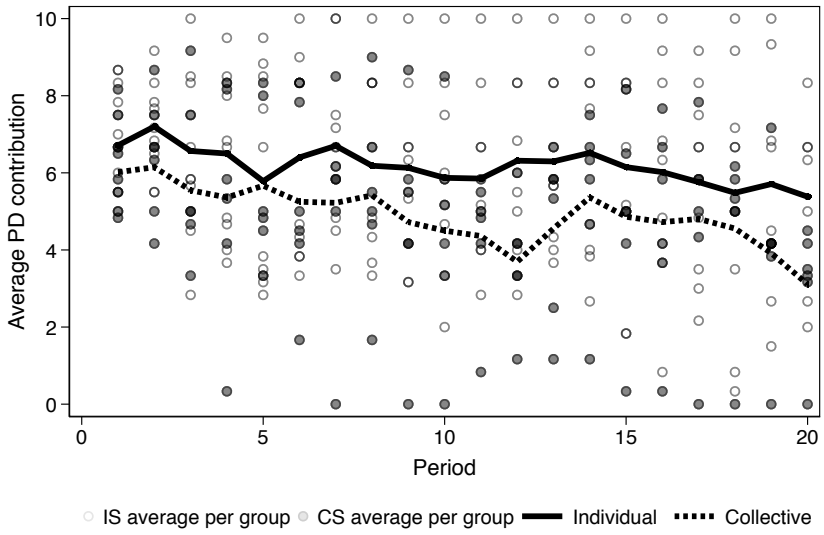


FIGURE 2.2: Average PD contributions per period

As speculated above, it thus seems to be case that intergroup collective sanctions can help to create a stable cooperative equilibrium, even though this is not as efficient as the one achieved under individual intergroup sanctions. It should be repeated, though, that intergroup individual sanctions are merely included here as a reference point. In real life, typically only collective sanctions are available.

With regard to punishment, Table 2.1 already showed that intergroup punishment rates were much higher in the IIS treatment as compared to ICS: 1.2 against 2.3 tokens – in line with theoretical expectations. Arguably, participants in the ICS treatment feared retaliation and therefore showed restraint. Or they were concerned with the unfairness of collective sanctions.

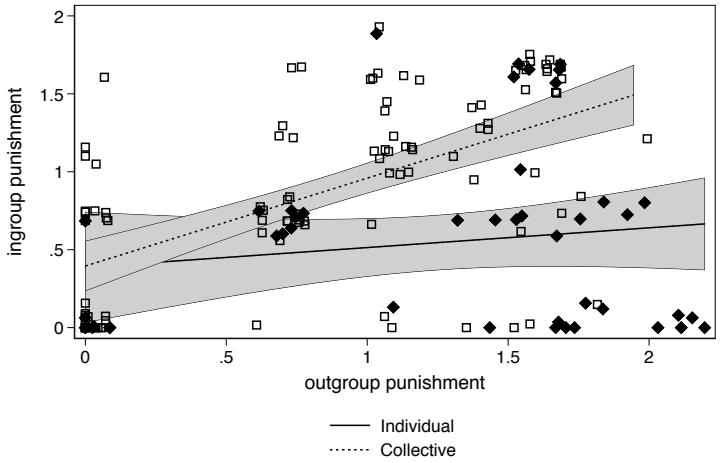


FIGURE 2.3: Ingroup leniency in individual intergroup sanctions

There are hardly any differences with regard to ingroup punishment, which was 1.7 tokens under IIS and 1.8 tokens under ICS. That is, there is no evidence that collective sanctions induce a higher degree of ingroup punishment. The fact that in IIS *intergroup* punishment is actually higher than *ingroup* punishment points towards ingroup bias: in making use of the availability of individualized sanctions, participants preferred to drive cooperation up by punishing out-group members rather than in-group members. This tendency is especially evident when we analyze separately only those punishment decisions where both the in- and out-group member were punished. As shown in Figure 2.3, we can see that under ICS, sanctions are applied more or less proportionally. In contrast, under IIS, there is clear evidence of in-group leniency, with in-group members receiving systematically less punishment than out-group members.

However, this conclusion about 'in-group leniency' should be taken with a significant grain of salt. The current experimental design does not allow to distinguish between several factors that can produce more symmetrical punishment reaction under collective sanctions. Under collective sanctions participants have an extra motivation to punish their own group members because they are unable to detect the specific targets in an out-group. The efficiency of outward punishment drops, making an in-group punishment more attractive option from purely rational point of view, whereas the hypothesis of 'in-group leniency' assumes that collective sanctions change an intrinsic in-group bias which has rather psychological grounds.

These impressions are confirmed if we look at punishment over time (Figure 2.4). In IIS, intergroup punishment exceeds punishment of ingroup members in all rounds up to the 17th. The decline in punishment rate in the last rounds is also typical for games with punishment stage.

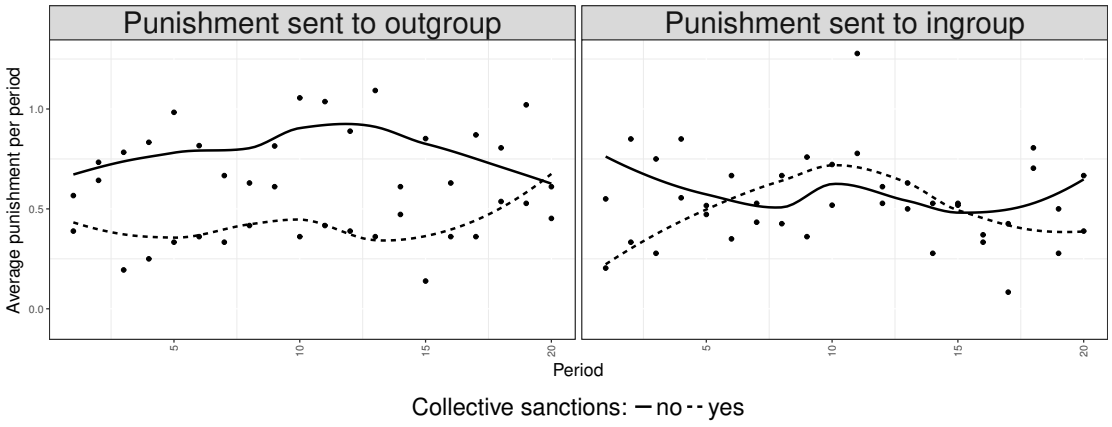


FIGURE 2.4: Average in- and out-group punishment sent per period in different treatments

Finally, we can find evidence for the relative higher *efficiency* of individual sanctions as compared to collective sanctions, which can further explain the higher cooperation rates under IIS as compared to ICS. Under IIS, there is a significant correlation between contributions and punishment . In contrast, under ICS, no clear such correlation is observable, which perhaps is

unsurprising given that punishment, from the perspective of the punished individuals, was applied somewhat randomly.

2.9 Discussion and conclusion

This paper began by exploring theoretically the efficacy of collective sanctions. Theoretically, an institutional regime in which an individual misdeed produces a risk of sanctioning for the entire group should result in a higher degree of cooperation between group members. Also in theory, this higher degree of cooperation should be achieved by means of increased ingroup third-party peer punishment of norm violators.

The logic behind this conjecture is simple: collective sanctions provide an additional motivation for ingroup members to find the person responsible for the punitive measures imposed on some or all members. Thus, knowing that a norm violation towards strangers will not easily go unpunished, when a potential perpetrator faces a social dilemma that involves a stranger, she will weigh more carefully whether it makes sense to behave antisocially. Members of the outside community become aware of the higher reliability of ingroup members, and this makes them more eager to cooperate in return. Therefore, theoretical works, like those by Stoff, 2006 and Fearon and Laitin, 1996, argue that collective sanctioning can explain why we observe lower rates of intergroup conflict than might be expected from work on ingroup bias.

My aim in this paper was to test empirically whether the introduction of collective sanctions has a beneficial effect on intergroup cooperation. Based on the result of the online study, the answer remains ambiguous. What I find is that intergroup collective sanctions clearly do not create a degree of cooperation that can be achieved under the 'gold standard', individual sanctions.

That said, the threat of intergroup collective punishment did help to stabilize intergroup cooperation rates. It may thus be concluded that collective sanctions can stabilize intergroup relations, but only at the price of relatively high welfare losses.

Although the current experimental design can be considered a productive starting point to inquire different sanctioning regimes in intergroup relations, it has one **serious** constraint. One of the fundamental reasons why groups may start using collective sanctions toward out-group members in the first place is information asymmetry that has been discussed in the introductory part of this chapter. But using the current design under individual sanctions the amount of information each participant has in the stage of third-party punishment when he or she observes the decisions of a randomly chosen pair, is the same: decisions of both in-group and out-group members are known, and in case of a punishment they will be sanctioned individually. That makes an *individual sanctions* treatment an 'ideal-type' benchmark used to compare with *collective sanctions* treatment, rather then a baseline of information asymmetry situation without collective sanctions. In order to check more rigorously whether collective sanctions indeed can produce beneficial effect on cooperation, in follow-up studies it will be necessary to compare it with an alternative 'baseline' **with** information asymmetry but **without** a possibility to punish out-group members.

Some other caveats to this conclusion are in order. As many other behavioral experiments, this study has limited external validity.[3] Inevitably, intergroup relations are presented in oversimplified form. For example, in this study, group members faced a binary choice to cooperate or to defect

[3]Lab experiments in particular are often considered to suffer from this shortcoming (Lucas, 2003), and this criticism is often applied *a fortiori* to online lab-experiments like the present one because of the relatively uncontrolled environment in which they take place. However, systematic evaluations show that there appears to be no significant differences between lab- and online experiments – see (Berinsky, Huber, and Lenz, 2012) for further discussion of the topic.

against outgroup members. In a more realistic context, instead of *either defecting or cooperating* with strangers, an individual can simply avoid outgroup members. Moreover, both in a real life and in some theoretical models (such as Fearon and Laitin 1996), only a minor share of a group needs to communicate and engage with strangers. The design that was used in this study does not account for these subtleties, but it is an important first step towards further understanding of mechanisms behind collective sanctions.

As a final note, I want to delineate aspects that could be added to future extensions of the existing design, and which might serve to further increase and deepen our knowledge of the subject. In this study, the only factors that united the groups were their communication with each other, the awareness of another group, and the ability to detect and punish defectors within their group. This common fate and history can be important factors in creating ingroup cohesion (Campbell, 1958), but may be insufficient for providing enough motivation to punish other group members. If ways could be found to further strengthen ingroup identities,[4] this may result in clearer results. Rather than strengthening group relations in the lab, another possibility would be to bring this line of study into the field. This would allow us to test how groups behave towards outsiders and among themselves when group cohesion and entitativity are not artificially induced in the lab, but provided by real life experience.

Finally, the present design also misses the important dimension of negative selection. Both in Stoff's and Fearon and Laitin's theoretical models, group members had the *option* to cooperate with either their own group members or with outsiders. This option attaches *signaling value* to the choice of cooperation partner, which is missing in the present design. The fact that

[4]For example, techniques from studies using the minimal group paradigm (Tajfel, 1982) could be used.

a person would like to cooperate with outsiders rather than ingroup members can be a signal that she has a bad reputation within the group, and thus has to look for contacts outside. This suspicion, in turn, can generate distrust among outsiders, and can be an additional reason why cooperation fails. This vicious circle can drive the situation to full conflict or to complete non-communication between groups.

Vice versa, if outgroup members are known for their higher degree of cooperation, at the extreme, ingroup members will *prefer* to cooperate with them. In those rare occasions the members of less cooperative group will try to tilt back the equilibrium to ingroup cooperation by punishing their own group members for norm violations, in order to restore their group's reputation. Allowing for the selective choice of cooperation partners may interact in intriguing ways with both collective sanctioning and peer punishment and, therefore, affect overall cooperation rates.

Chapter 3

Crime, peer punishment and collective sanctions: a lab experiment

Abstract

This paper analyzes how the introduction of collective sanctions affects the willingness to punish norm violations. The chapter presents a lab experiment in which participants can choose to take money from a charity. After having taken their own decision, they can observe the decisions of others, and can decide to punish them. I test whether collective sanctions reduce the rate at which people take from the charity, and at which they punish their peers under a) an individual sanctions regime, or b) a collective sanctions regime. The results demonstrated that collective sanctions significantly increased the frequency of peer punishment. However, this increased rate of punishment did not go along with lower crime rates. I conclude that, compared to individual sanctions, collective sanctions are inefficient.

3.1 Introduction

The aim of this chapter is to assess whether collective sanctions affect an individual's propensity to make ethically questionable decisions. For brevity, I will call decisions of this kind crimes. We may define a 'crime' as a risky action that can produce individual benefits to the actor, while at the same time bringing about harm towards a third party. In the experimental literature that seeks to study crime in the lab, the amount of harm caused is usually modeled to exceed the benefit for the perpetrator. This is done in order to distinguish 'crime' from situations of pure wealth redistribution (Falk and Fischbacher, 2002).

We speak of collective sanctions when an external authority (either an individual or an entire group) makes the decision to punish either an entire group, or a random member of a group. Both types of punishment are common in experimental investigations (e.g. Dickson, 2007 and Fatas, Morales, and Ubeda, 2010). To understand why random punishment can count as 'collective', consider that the randomness of punishment produces the same results as the punishment of an entire collective, assuming that objects of punishment are risk-neutral: "Sanctions are collective when they are *threatened* against or imposed groups of two or more individuals" (emphasis added). That is, "collective" should be understood in *ex ante* sense, to take account of the fact that sanctions directed against a single group member chosen at random will have the same expected disutility for group members as sanctions divided evenly among all group members (Levinson, 2003, p.277).

As discussed below, collective sanctions may affect crime rates in many different ways. Here I will focus on one specific mechanism: the effect collective sanctions may have on peer punishment. As discussed in Chapter 1, collective sanctions hold the potential to solve the information asymmetry between those interested in preventing crime (the outgroup) and potential

criminals (the ingroup). Collective sanctions provide an incentive for a group of potential criminals to monitor and punish each other. This threat of horizontal punishment from their peers will, in theory, avert aspiring criminals from actually committing a crime.

3.1.1 Background: use of collective sanctions in response to crime

In the real world, collective punishment is used in response to crime or crime-like offenses chiefly in three areas: offences committed in the context of interethnic relations, breaches of corporate responsibility, and the management of teams.

The situation in which members of one ethnic group punish random members of another ethnic group for a crime committed by a single member is a well-known mechanism to curb crime levels in the situations in which the central authority is weak or non-existent (see also Chapter 2). In parts of Tanzania, the vengeance for adultery spilled over the loving triangle of lover, cheating wife and cuckold, making the entire village of the offender responsible: "...if the injured husband did not find the adulterer he might kill any village-mate of his enemy. Such an attack commonly led to war between the two villages." (Wilson 1987 cited by Nakao and Chai 2011). In this example, collective punishment is not a reconciliatory or deferring measure, but a direct way into full-scale conflict. That is why utilitarian individualists such as Hardin warn us about the dangers of aligning group and individual interests via the mechanism of collective responsibility (Hardin, 1995).

But despite – or in fact because – of this danger, collective sanctions may also *stabilize* intergroup relations. The threat of violence spiralling out of control may deter defections and thereby support intergroup cooperation (Fearon and Laitin, 1996). One interesting feature of such retaliation against

the entire group is that, in theory, it should produce less defection and more cooperation even if the group size is small, such as in a three player Prisoner's Dilemma (Stoff, 2006).

Corporate and vicarious (third-party) responsibility is the second large area where collective responsibility for individual offenses is widely applied. Companies often are often held responsible for the actions of their employees. Examples include the withdrawal of the license of a restaurant when an employee fails to meet sanitary standards, the shutdown of a liquor store where a clerk sells alcohol to a minor, or the closure of a brokerage by The United States Securities and Exchange Commission when a single broker is found guilty of inside trading, to name just a few.

A recent example is the fate of the auditors from the Arthur Anderson auditing company: after being involved in the Enron affair, the company was heavily fined. And even though just a handful of auditors were personally involved into the fraud, "the sentencing of the corporate entity ruined the livelihoods of all the other Arthur Andersen partners in the world whom had no involvement whatsoever with the scandals" (Rönnegard, 2015, p.68).

A different example comes from Pettit, 2007. Pettit starts his paper on corporate responsibility using an example of a ferry accident in the English Channel: due to the negligence of the managing company about two hundred people drowned. He argues that in a case like this "it can make good sense to hold that while the individuals involved may not bear a high degree of personal responsibility together as a corporate enterprise they should carry full responsibility for what occurred" (Pettit, 2007, p.171). Transferring individual responsibility upon the entire company and making all its stakeholders pay for its faults follows a consequentialist logic. Otherwise, "[i]f employers can externalize liability costs through shallow-pocketed employees, their level of precaution-taking will be inefficiently low" (Levinson, 2003).

The third area of wide-spread usage of collective sanctions is the management of teams. Team managers face informational disparity when they have to monitor the behavior of their subordinates. When direct supervision is impossible or too costly, managers rely on collective sanctions, "punishing all work group members for one individual's misconduct" in the hope that this would make peer reporting more acceptable (Trevino and Victor, 1992). A survey conducted in the year 2000 among corporate employees in the US revealed that about one third had witnessed some kind of misbehavior in their work during the previous year (Joseph, 2003). Collective sanctions can help to stop team members from misconducting as their colleagues can often actually observe the crimes committed by their colleagues. They have thus a significant advantage over their team manager, who is often seated separately. Often, the groups that have to share collective responsibility for a misdeed of one of their members are not institutionalized, but temporary. This is the case for teams of scholars writing an academic paper, for instance: in the case of plagiarism made by one scholar the reputation of everyone in a group is ruined. Such collective responsibility sometimes is set in legal terms, too. Stanford University's policy on multiauthored research papers assigns "[a]ll authors in a group effort... shared responsibility for the published results" (Levinson, 2003, p.141).

3.2 Why collective sanctions may work

How effective are sanctions in general for deterring criminal behavior? Rationalists, usually associated with Hobbes and Locke, claim that people react positively to rewards and negatively to sanctions, and change their behavior accordingly. Thus, any punishment for antisocial behavior produces a prosocial effect. As Locke states "Good and evil, reward and punishment are the only motives to a rational creature: these are the spur and reins whereby all

mankind are set on work and guided" (Locke, 2007). A recent re-incarnation of this principle was formulated by Hardin, 1968, who argues that the only effective tool to curb self-interest for the sake of public good "mutual coercion mutually agreed upon".

Sanctions are usually thought of deterring both a wrongdoer and his or her potential followers from committing a crime. This deterrence is twofold: general and specific (Travis, Western, and Redburn, 2014). General deterrence preemptively repels an individual from crime commitment via the punishment threat. Specific deterrence makes a person who experienced a sanction more reluctant to repeat this action. The impact of collective sanctions on these two dimensions of deterrence is mixed. On the one hand, under collective sanctions there is a higher probability of suffering from the sanctions even if a subject has no intentions to commit a crime, so the introduction of collective sanctions makes general deterrence less convincing. On the other hand, the number of those who personally experience such sanctions will be larger, so specific deterrence under collective sanctions may seem to be more efficient.

Collective sanctions usually make sense when there is a disparity of monitoring costs between group insiders and outsiders. This disparity can be either informational or executional. In case of informational disparity, the detection of a non-cooperative type is more costly or difficult for an outsider than for an insider. A police officer may not know the name of the local pusher while those who live in the block know the identity of the dealer. The executional disparity appears if the cost of punishment or prevention of certain type of behavior is lower for those inside the group as compared to outsiders. For example, such an extra cost of punishment can be a result of limited jurisdiction: is a person who has an interest in sanctioning deviant behavior eligible to do it? It is perceived legitimate for employers to sanction employees, but the same is inappropriate for clients. Similarly, parents

can punish their children if these misbehave towards strangers. However, if the strangers themselves take the matters to their own hands, it may be a very costly enterprise for them. The reputational damage that the family incurs for not punishing their own misbehaving member is one of the possible mechanisms to correct this disparity: without putting the question of parental jurisdiction in the question, it gives outsiders the leverage to guarantee that punishment will be imposed via the "internal middlemen", the parents.

From these two types of disparity between the authority (the external agent) and ingroup members follow three different logics that may drive the authority's intention to introduce collective sanctions as a reaction on a crime (Nakao and Chai, 2011): informational, preferential and functional.[1] According to the informational logic, the authority punishes a random suspect in a culprit's social group because of an informational disparity: the outsiders do not know the identity of the wrongdoer. As Hardin states, "groups are apt to have better information about their members' actions than about the actions of people in other groups." (Hardin, 1995, p.118). According to the preferential logic, collective sanctions are used because the external authority expects that punishing innocent in-group members would bring more harm for a wrongdoer than a direct punishment suffered by him- or herself. And lastly, the functional logic holds that group members will deal with the perpetrator more effectively. In the functional logic, peer punishment is the main mechanism through which collective sanctions translate into intra-group discipline.

3.3 Why collective sanctions may fail

From the descriptions above it may seem that there is a little doubt that – through the incentivization of peer punishment – collective sanctions will

[1] These motives have been discussed in more general terms in Chapter 1.

be an effective tool for crime deterrence. However, there are good reasons why we should be doubtful that collective sanctions are a universal cure for crimes and other misdeeds.

For a start, the rational approach to sanctions – that sanctions will always work to move a subject in the desired direction – has been attacked on both theoretical and empirical grounds. Criminal propensity theory suspects that punishment is a bad tool to deter crime because criminals tend to be individuals who are criminally-prone by nature. Being impulsive, high risk-taking and present-oriented, criminals are also the least likely to be deterred (Wright et al., 2004). Even worse, sanctions may not only fail to prevent crime, but may crowd out prosocial behavior. Starting from the 1970s, psychologists have explored how rewards and punishment may in fact *undermine* the intrinsic motivation for some socially beneficial actions (Kruglanski, Friedman, and Zeevi, 1971). The presence of a sanctioning system undermines the belief of participants that a collective good can be produced without them (Mulder et al., 2005).

Collective sanctions may also fail because their internal logic harms basic principles of just sanctions. These principles are the following: 1) Sanctions should be fair: there should be a causal chain between punishment and the wrongdoing; 2) Sanctions should be proportional: there should be correlation between their size and the size of the harm the wrongdoing produced; 3) Sanctions should be bounded, or well-defined: for the specific crime, the size and duration of sanctions should be known beforehand and finite. Collective sanctions do not perform well on all three dimensions. By definition, the causal chain is broken between action and counter-action: under collective sanctions formally non guilty actors bear the burden of punishment – just because they belong to the same group as a wrongdoer. Depending on the specific rules of punishment implementation, collective sanctions may also fail on proportionality. For example, collective sanctions could be additive,

meaning that if two perpetrators are caught, this cause twice the amount of sanctions being imposed on the group as if one perpetrator was caught. With such a rule, if the crime rate is high, the amount of punishment received by each individual member will be disproportionally high. Finally, collective sanctions also lack clear boundaries: if the crime rate is not known, the amount of sanctioning that an individual agent will receive is also unknown.

Another problem with collective sanctions is that they are perceived as procedurally unfair, simply because the entire group is punished for a misdeed and there is no procedure that determines who specifically is guilty. To be effective, punishment should be applied in a situation where the object of punishment feels guilt, otherwise punishment fails to be an effective deterrent or even provokes counter-punishment (Hopfensitz and Reuben, 2009). Procedural fairness is crucial in raising individual satisfaction with the outcomes. This principle has also been shown in the lab. Even if personally, subjects suffer from a specific game outcome, they are less upset if they find the process that resulted in this outcome procedurally fair (Lind and Tyler, 1988; Culnan and Armstrong, 1999).

A failure to meet the requirements of fairness and proportionality undermines the legitimacy of a regime, making it less effective. Any sanctioning regime consists of an asymmetric dyadic relation where the active party, the administrator of sanctions, produces some harmful actions upon the passive recipients. If the sanction is perceived as unfair, however, the recipients of the sanction may rebel and cease to cooperate to signal their disapproval. Indeed, their willingness to send such a signal may be so strong that they may be ready to incur high costs to do so.

The use of collective sanctions may also be hampered by a reluctance by the active side in the sanctioning dyad, the administrator, to actually use them. The administrator may refrain from imposing sanctions when their fairness is doubtful for several reasons: moral costs, the expected threat of

retribution, and empathy. First, there are usually moral costs associated with controlling and sanctioning others. The questionable legitimacy of sanctions increases these costs. Second, a person applying supposedly unfair sanctions may expect retaliation from the recipient. Third, the administrator may put himself in a recipient's shoes: if he would end up in a similar situation, he would prefer not to be the object of similar sanctions.

3.4 How do collective sanctions affect peer punishment?

Collective sanctions are defined as a negative measure that is applied to a group 'from above' – either by a centralized authority, or by another group. Conversely, peer punishment is a horizontal way of inflicting harm on one's peers. Peer-punishment is usually used if an actor believes that a group-norm has been violated, or as a retaliation in response to a previous act of punishment. By imposing collective sanctions upon the entire group, the authority or external group may push the target group to establish norms in a specific area and to provide enforcement of this norm. This outside inducement is important, since group do not establish or enforce group norms about every conceivable situation. Instead, "[n]orms are formed and enforced only with respect to behaviors that have some significance for the group." (Feldman, 1984, p.47). The threat that the misbehavior of a single member will have everyone in a group be sanctioned provides such a ground for internal norm enforcement.

There are competing explanations to explain peer-punishing behavior. In Fehr and Gaechter's experiment on altruistic punishment (2002), people were expected to take part in the creation of a public good, and then, after observing the contributions of their group members, could, at a cost to themselves,

send deduction tokens intended to decrease their peers' final payoffs. The authors link the propensity to engage in peer punishment to the idea of fairness. Indeed, there is a lot of evidences that one of the main motives to punish less cooperative members of the community is inequality aversion and retaliation for unfair actions (Falk, Fehr, and Fischbacher, 2005). In contrast, Casari and Luini, 2012 theorizes that individuals have a hard-wired *taste* for punishment, and engage in punishment even without taking into consideration instrumental or fairness considerations. Applied to a situation where peer-sanctions are combined with collective sanctions, both approaches predict higher peer-punishment under collective sanctions. Observing a peer earning an additional income from a deviant behavior (here, free-riding in the public good game) *and* suffering *additionally* from a collective sanction being imposed in response to that behavior, peers have an extra motivation to retaliate against their deviant peer. Collective sanctions and peer punishment may therefore be considered two parts of the "ethical infrastructure" of a group (Tenbrunsel, Smith-Crowe, and Umphress, 2003). Formal rules imposed from above – the collective sanctions – are combined with a set of informal rules enforced by the collective itself.

However, even though the two sets of rules may belong to the same ethical infrastructure, this does not mean that there cannot be any conflicts between them. Quite on the contrary, the norms imposed through these two channels can be totally different, because the objectives of external and internal sanctioning actors often oppose each other. Attitudes to whistleblowing within a company may serve an example of such a conflict: while the management of an organization is interested in whistleblowing, the unofficial sanctioning system usually severely punishes snitches. Tenbrusel et al. cite the story of a petty employee theft at the Wisconsin mill in Wisconsin. After blowing the whistle, the informant was found dead "at the bottom of a 20-foot holding vat for tissue pulp. A jump rope attached to a 40-pound weight

was tied to his neck" (Worthington, Chicago Tribune, October 12, 1993 cited by Tenbrunsel, Smith-Crowe, and Umphress 2003). There is a direct conflict between two spheres: "At the heart of this story are two contrary views on the ethical principle related to employee theft: the formal perspective, which classifies such behavior as unethical and illegal, and the informal principle that employee theft is an action that should be tolerated and, perhaps more important, not 'snitched on' by a fellow union member." (Ibid). Collective sanctions can be seen as a mechanism which aligns the interests of out- and in-group members, thus solving this conflict. Stated in such terms, the main research question of this study is to investigate how institutional changes in the formal sphere – the introduction of collective sanctions – affect the informal sphere – the level and use of peer-punishment –, or, in short, the ethical infrastructure of the group.

3.5 Hypotheses

We can thus formulate the following hypotheses, to be tested in the experimental investigation below:

Hypothesis 1: Under a collective sanctions regime we will observe more peer punishment.

The prospect of being punished for doing nothing will push honest actors towards the peer punishment option. This is not the case under individual sanctions, where a person is sanctioned by an external authority for his or her own actions only. In the case of individual sanctions, the motivation for peer punishment thus is purely altruistic. In contrast, since in the long run, peer punishment under collective sanctions reduces the chance of being punished by the external authority for each member of the group, members have a self-interest in punishing.

Peer punishment under collective sanctions regime serves as a substitute to external sanctions. That is, at least the logic behind its implementation by policy makers when outside detection of an actor responsible for a misdeed is unattainable or overly costly. One of the most characteristic example of collective sanctions that we can encounter in real life is a conflict between a teacher and a school class. For instance a teacher trying to detect who is responsible for a broken window, grounds the entire class. If group members have a choice to *preventively* punish the offender at the prospect of potential external sanctions, they are more eager to do it, especially if they are risk-averse (it should be noted here, that the current design does not include risk aversion measures, so this specific assumption will not be tested).

Hypothesis 2: Under a collective sanctions regime, the increased risk of being punished by one's own peers decreases one's propensity to commit crimes/deviant acts.

There are two mechanisms that can drive the incidence of crimes when collective sanctions are introduced, which work in opposite directions. On the one hand, the increased incentives for peer punishment increase the cost of committing a crime, making it a less attractive option. On the other hand, the fact that you can be punished by an external authority anyway, notwithstanding your own decisions, may make crimes more attractive: if you are likely to pay for what you have not done, it makes sense, at least, to receive the extra payoff associated with the crime to compensate for the punishment. It is hard to predict which mechanism prevails, but if the prevalence of peer punishment substantially increases due to collective sanctions, it is most likely that on balance, participants will be discouraged from choosing the criminal option.

3.6 Experiment

The present section puts the two hypotheses above to an empirical test. To this end, I designed a dedicated behavioral game, which I then tested by means of a lab experiment. I first introduce the formal logic of the game, then introduce the experimental procedures, and outline the results.

3.6.1 Formal game description

Let's consider the situation of two team members, both of whom have the option to commit a crime (embezzle a company's money, take a bribe etc.) Each of them simultaneously decides whether to be honest and earn a *regular* salary, w, or be dishonest, and seek a higher income Y_c, which, however, inflicts a negative externality on the other team member. This option is referred to as the criminal option or crime, and the team member who commits the selfish act as the perpetrator.

If only one team member chooses to commit the crime, then the other one has a choice to punish his or her partner. Punishment costs c for the punisher, and $F > c$ to the person who is punished. If the harmed person decides to stay idle, then, with a probability of p, the criminal decision will be checked. In this case, the perpetrator will pay a fine of **F**, and the other player will be fined f, which is the fine for *not punishing* his or her criminal peer.

In the experiment, as in the other chapters, there will be two treatment conditions: an individual sanctions regime, and a collective sanctions regime. For the individual sanctions regime, $f = 0$, and for the collective sanctions regime, we will set $f = F$.

Game-theoretical predictions are rather straightforward. If the cost of punishing a partner exceeds the expected cost of collective punishment ($c > pf$), then a participant will always stay idle. Then, knowing that there is no chance of being punished by his peer, the second player will commit the

crime if he knows that the net profit of the crime is positive ($Y_c > pF$). So if both participants have similar estimations of their potential profits and the probability of being caught, and there is complete information (that is, both participants know about each others' preferences and estimated probabilities), there will be two pure equilibria: both will either commit the crime or none will. If information is incomplete, peer punishment becomes possible. Depending on whether players under- or over-estimate their partner's costs, peer punishment may or may not occur.

3.6.2 Experimental design

The following section introduces the experimental procedures. I designed a two-player experimental game in which I operationalize crime in terms of withheld donations to a charity. The game was played in two treatment conditions, an indvidual sanctions treatment, and a collective sanctions treatment.

The experiment was conducted with 160 participants at the Indian Institute of Management in Ahmedabad (IIMA), India, during five sessions that took place between January 29th and February 13th 2015. The experimental script and more detailed information on the session procedures are provided in the appendix to this chapter.

The experimental setup largely followed the theoretical outline above, and all design choices are summed up in Table 3.1 below. Upon arrival, people were matched with a partner to form a team of two. They then took their decisions individually.

The decision task consisted of a simple choice, which was repeated over several rounds: to take a blue envelope, or, instead, take a green envelope. Taking an envelope simulated either taking a socially approved and socially beneficial, 'regular' course of action (represented by the blue envelope), or

instead choosing a 'deviant', or 'criminal' course of action (represented by a green envelope).[2]

There were ten green and ten blue envelopes to choose from. Each blue envelope guaranteed an income of 10 rupees for a participant. In addition, upon the choice of the blue envelope, 25 rupees would be donated to a charity. The choice of a green envelope, in contrast, typically guaranteed a payoff of 25 rupees for the participant, but reduced the donation to the charity to 0.

If we assume that the trust to the effectiveness of NGOs is high enough the socially optimal behavior is to choose blue envelopes. In order to reduce the potential bias resulted by a choice of a specific charity, we announced the name of charity after the study has been completed. Choosing green envelopes was individually profitable, but collectively harmful, resulting in an external negative externality of 15 rupees. Choosing the green envelope therefore corresponds to selfish and deviant behavior – henceforward referred to as the 'crime' in this text.[3]

Yet there was another twist: out of the 10 green envelopes, 3 actually were empty. Choosing a green envelope, the participants thus had a 3/10 chance of ending up with no payoff. This danger of drawing a blank envelope was the implicit punishment for committing the 'crime'. Note that despite this potential for punishment, the individual expected income from taking the criminal option nevertheless remains higher than the expected income from the 'regular' option (taking the blue envelope) – 17.5 rupees for the crime vs. 10 rupees for the 'regular' behavior. In order to test whether people behave

[2]In the instructions or during the games, at no point loaded terms such as 'punishment', 'crime' or 'delinquency' were used (see the Appendix for the text of the instructions). I here use these terms for the sake of referential homogeneity with the theoretical model.

[3]It should be mentioned that in most peer-punishment studies, participants are typically first faced with a social dilemma, not a donation task. In a social dilemma, an individual profits from selfishness unless everyone chooses the selfish alternative, in which case the whole group loses (Schroeder, 1995). These social dilemmas are usually modelled as public goods games. In the current design I avoid the social dilemma situation as a baseline for peer punishment and collective sanctions. I do so in order to disentangle the personal motive of being angry with a free-rider from the motive of retaliation against the moral unfairness of the perpetrator's actions.

differently in response to changing risks of discovery (cp. Loewenstein et al., 2001), in a variation of the experimental setup, the chance of detection for perpetrators was set to $p = 0.1$ rather than $p = 0.3$. That is, in this variant, only 1/10 of the green envelopes was a blank.

TABLE 3.1: Design

	Variables	Individual Sanctions	Collective Sanctions
Y_h	Payoff for honest behavior (blue envelope)	10	
Y_d	Payoff for deviant behavior (green envelope)	25	
NE	Negative Externality	15	
p	Probability to get 0 (blank envelope) in case of criminal behavior	10%; 30%	
w	Cost of peer punishment	5	
c	Fine for Collective Sanctions	10	0

The two treatments conditions differed with regard to the consequence of drawing a blank envelope. In the individual sanctions treatment, if a person chose an empty green envelope, this did not affect the payoff of the other team member. In contrast, in the collective sanctions treatment, if an envelope was empty, the other team member's payoff was diminished by 10 rupees, meaning that both the perpetrator and his or her team member would finish the round with 0 rupees. Therefore, in the individual sanctions treatment, only perpetrators got punished (with a certain probability). In contrast, in the collective sanctions treatment, perpetrators got punished for committing the crime *and* the other member of the team was penalized for staying idle.

In addition to this 'institutional' punishment by an external authority (the game designer), participants could engage in costly peer punishment. This was implemented as follows. If a team member chose the green envelope, the other member was given the option to pay 5 rupees to punish his or her peer. Punishment immediately set the payoff of the criminal team member to 0 – rather than the likely 25 rupees from the green envelope. That is, in a

round were one team member chose the green envelope, and the other chose the blue envelope *and* decided to punish his or her peer, the former would end the round with 0 rupees, and the latter with 5 rupees. In situations where both team members chose the green envelope, both could punish each other. If, instead, both team members chose the 'honest' course of behavior and chose the blue envelope, peer punishment was unavailable.

This latter choice was made to resemble real-life situations. In real life, when there is no crime, there is no punishment. So if nobody chooses the green envelope, the game would end here. In order not to lose data in these situations where none of the team members committed the crime, participants took their punishment decisions using the strategy method. That is, they indicated how they would behave, both if the other team member had committed the crime, or if s/he had not.[4] The real action of their team member was revealed at the end of each round. The game thus proceeded in three stages:

- **Stage I:** Participants take their decisions what kind of envelope to choose.
- **Stage II:** The participants take their decision on costly peer punishment using the strategy method.
- **Stage III:** Participants draw one envelope out of 10 (either green or blue) and get rewarded according to the rules of the game.

In cases with full peer punishment in the second stage, i.e. if all members who chose the green envelope were punished by their peer, the game ended after the second round as there was no more money to be taken away.

[4]The strategy method has been shown to be as reliable as the direct-response approach (Brandts and Charness, 2011).

I tested the effect of individual vs. collective sanctions on the participants' behavior using both within- and between subject designs.[5] 80 of the 160 participants played the game in the between-subject design: 44 under the individual sanctions regime, and 36 under the collective sanctions regime. The other 80 participants played in the within-subject design. That is, they played under both individual sanctions and collective sanctions. To avoid order-effects, half of the participants first faced the individual sanctions regime and then the collective sanctions regime, while the other half faced the reverse order. The reduced probability of detection – $p = 0.1$ instead of $p = 0.3$ was implemented for 58 participants in the within-subject design. These treatment conditions are summed up in Table 3.2.

TABLE 3.2: Overview of treatment conditions

		Individual Sanctions	**Collective Sanctions**
Between-subject design	$p = 0.3$	44	36
Within-Subject Design	$p = 0.1$	58	
(IP-CP; CP-IP*)	$p = 0.3$	22	

*Half of each within-subject subgroups played "Individual Sanctions" first, and half played "Collective Sanctions" first.

3.7 Results

The two hypotheses to be tested were that 1) under collective sanctions there would be more peer punishment, and 2) that under collective sanctions there would be less crime. I present the results for the within- and the between-subject treatment separately, starting with the results of the within-person design, summed up in Table 3.3.

[5] Both approaches have their advantages and disadvantages (cp. Bellemare, Bissonnette, and Kröger, 2014), but without prior experience, it is not possible to predict which specific design will bring about more robust results.

As for peer sanctioning, punishment rates appear to be similar across conditions. Under the baseline rate of detection of $p = 0.3$, the number of those who punished their peers was 24.4% in IS, and 27% in CS. With the lower detection probability of $p = 0.1$, these values changed to 25.1% in IS and 32.7% in CS. While these results seem to support hypothesis 1 that there would be more peer-punishment under collective sanctions, differences are not statistically significant.

With regard to the crime rate, i.e. the share of participants choosing the green envelope, the following pattern is observable: under IS, 36.3% chose a green envelope, while under CS, the rates of perpetrators was slightly lower (30% chose the green envelope). While on the face of it, this result seems to support hypothesis 2 that crime rates should be lower under CS, this difference in behavior is not statistically significant. When the probability of sanctions was decreased to $p = 0.1$, the number of those who chose the green envelope grew strongly to 61% in IS, and 57% in CS – a result which conforms with the increased expected utility of taking the green envelope. However, again the difference between conditions is not statistically significant.

TABLE 3.3: Results overview within-person design

Treatment	Individual Sanctions (IS)		Collective Sanctions (CS)	
Probability of detection	0.3	0.1	0.3	0.1
Crime rate (% choosing green envelope)	36.6%	61%	30%	57%
Punishment (% punishing their peers)	24.4%	25.1%	27%	32.7%

Looking at the dynamics of crime and peer punishment across periods, the overall impression of similar behavior under IS vs. CS seems to be confirmed. In the first three rounds there are strong differences between treatments both in 'crime rate' and peer punishment frequency. Later on, this

difference fades away. The main explanation for this phenomenon is the 'learning effect': it is rather typical to have a certain amount of noise at the beginning of the multi-round game, which disappears as participants start understanding the mechanics of the game and an incentives structure better.

While IS seems to produce slightly higher crime rates and lower peer punishment, confidence intervals overlap at most times, meaning that differences between the two conditions fail to reach statistical significance.

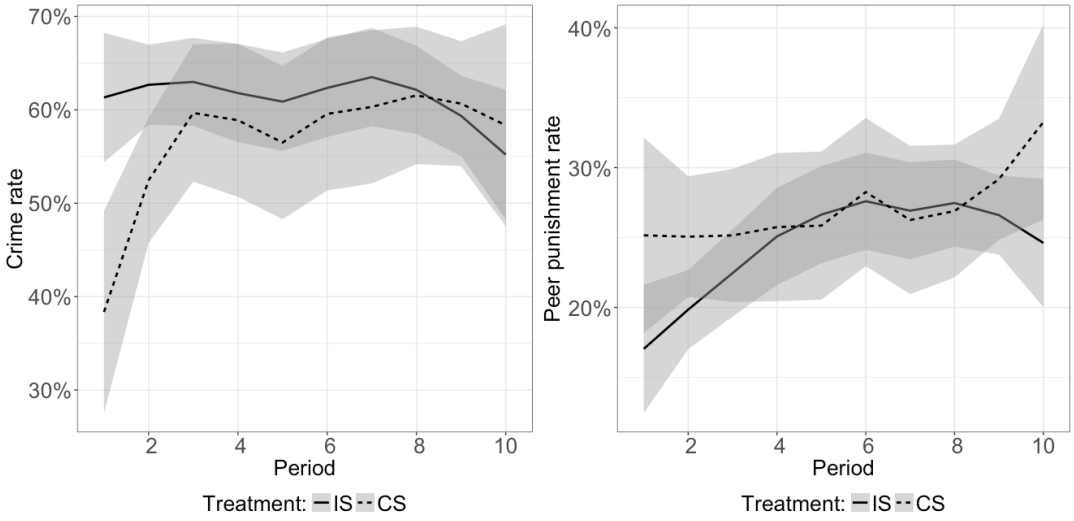


FIGURE 3.1: Crime rate and peer punishment over time in the within-person design

More intriguing results are obtained in the between-subject setting, summarized in Table 3.4. With regard to the overall crime rate, the pattern we just observed seems to be reversed, an impression that is confirmed in Figure 3.2.

TABLE 3.4: Results overview between-person design

Treatment	Individual Sanctions (IS)	Collective Sanctions (CS)
Crime rate (% choosing green envelope)	58%	62%
Punishment (% punishing their peers)	21%	31%

Now, the crime rate is higher in CS (at 62%) than in IS (at 58%), contrary

to the hypothesis. However, once again, differences are not statistically significant. It therefore appears that, in fact, IS and CS produce very similar crime rates. Put another way, individual sanctions and collective sanctions seem similarly effective in containing or not containing crime.

When we look at peer-sanctioning behavior, however, we observe that here the differences between CS and IS are substantial and statistically significant. The peer punishment rate is 21% under IS, but a whooping 31% under CS. As shown in Figure 3.3, the peer punishment is higher in CS than IS in the majority of rounds, and confidence intervals hardly overlap.

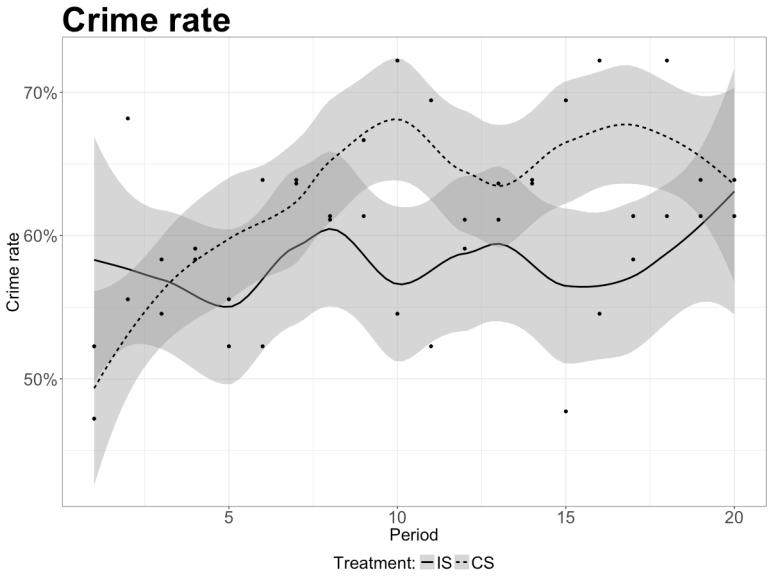


FIGURE 3.2: Crime rate (share of participants choosing the green envelope) over time under IS and CS

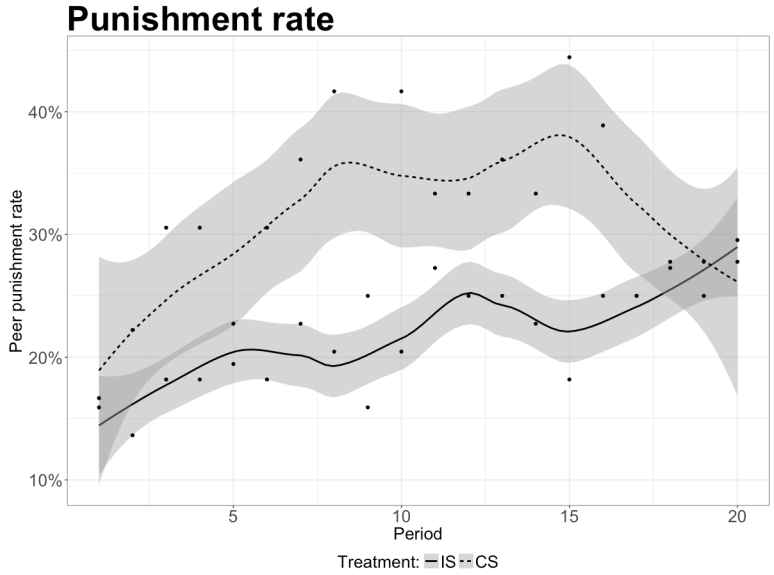


FIGURE 3.3: Punishment rate (share of participants choosing to take the criminal income of their peer away) over time under IS and Cs

The reasons for this difference is illuminated if we look at the *kind of decisions the punishers took at the first stage* (Figure 3.4). Looking first at participants who took the 'honest' or 'regular' decision by taking the blue envelope, we can see that there are hardly any differences between the two conditions. Under CS, the rate of those who chose a blue envelope and punished (34.1%) is almost exactly the same as under IS (34.6%).

If participants took chose a green envelope themselves, however, then, under IS, only 12% of them chose to punish those who also chose the green envelope. That is, under IS, criminals do not punish other criminals.[6] In contrast, under CS, the number of those who also took green envelopes themselves is 29% – a 2.4 times higher rate! So under CS, both 'regular citizens' and criminals punish – or retaliate against – other criminals.

[6]The 12% who do so anyway may be motivated by 'spiteful' motives in the sense of Jensen (2010).

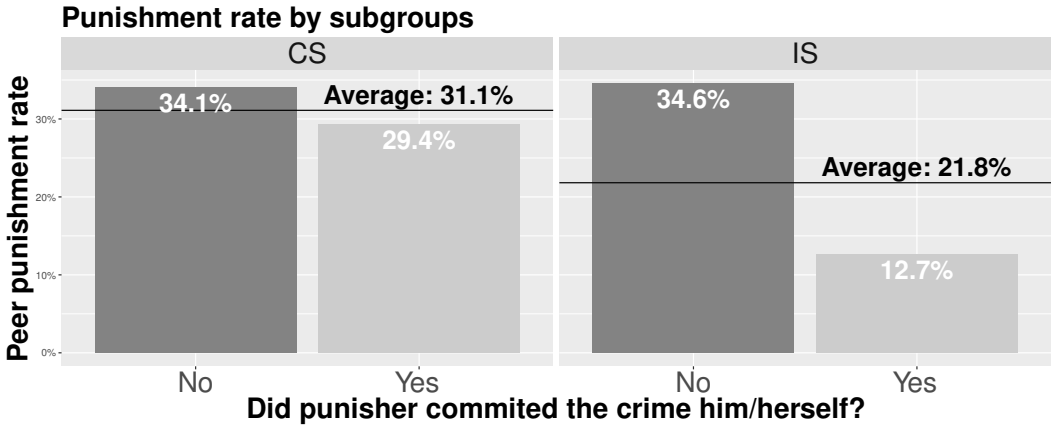


FIGURE 3.4: Punishment behavior by perpetrators of crimes vs. non-perpetrators under IS and CS

I now use a multivariate model to assess whether peer sanctions are actually effective to curtail crime under the different punishment regimes. In the model, I consider how the likelihood of choosing the green envelope, i.e. committing the crime, is affected by a) the choice of envelope in the previous round, and b) peer punishment in the previous round. I also control for gender and self-reported trust to improve precision. The dependent variable takes the value 1 when a green envelope is chosen, and 0 if a blue envelope is chosen. For the analysis, I use a probit model with random effects for the rounds of the experiment, clustering standard errors at the level of the individual.

Table 3.5 reports the results. We can see that, as already shown above, the baseline propensity of committing a crime is subtler higher (4% difference) under collective sanctions, but that this difference is not statistically significant. The choice of envelope in the previous round has little predictive power for the choice of a green envelope in the subsequent round, and neither have gender or self-reported trust.

Although not statistically significant, it is nevertheless interesting that the choice of a green envelope in the previous round (slightly) *decreases* the chance that a participant will choose the same envelope at the next stage.

There are two different explanations: either people believe they will not be lucky twice, and thus prefer to choose the blue envelope after choosing the green one, fearing that they might otherwise pick an empty envelope. Or they prefer to *vary* blue and green envelopes as a gesture of sharing with the charity at least sometimes.

The most interesting results pertain to the effect of peer punishment. Here we can see that under IS (represented by the constitutive term 'Peer-punishment at $t-1$'), peer punishment has a slightly discouraging effect on committing the crime in the subsequent round. Under CS (represented by the interaction term), however, this effect is much stronger and highly statistically significant. This means that under CS, peer punishment is actually *more* effective than under IS in reducing crime.

TABLE 3.5: Probit random effects model of criminal behavior on previous punishment and other covariates

	Choice of green envelope at $t = 0$
Col.sanctions	0.538
	(1.57)
Peer-punishment at $t - 1$	-0.0928
	(-0.71)
Col.sanctions X Peer-punishment at $t - 1$	-0.571**
	(-2.96)
Choice green envelope at $t - 1$	-0.0621
	(-0.62)
Gender	0.516
	(1.37)
Trust	0.344
	(1.00)
N	1,520

t statistics in parentheses

$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

It therefore may seem paradoxical that crime rates appear to be the same under IS and CS. We can speculate that this seeming paradox is due to an effect mentioned earlier: the very fact that under collective sanctions they might be punished without being guilty might actually push people to behave more deviantly under this sanction regime – resulting in a higher baseline rate of crime under CS as compared to IS.

This means that the similar crimes rates in IS and CS come that the price of efficiency. Indeed, due to the higher occurrence of punishment under CS,

participants 'burned' a lot of money, and ended up with lower average earnings per round than under IS: 9.5 rupees vs. 11.8 rupees. The fact that peer-punishment is used more excessively under CS also showed in answers to a survey, delivered after the experiments. Here I asked participants whether they would be involved in peer punishment more if the price of such a punishment would decrease from 5 to 3 rupees. Under IS, the majority of the participants (61%) claimed that they would *not* punish more often in response to the decreased price. The opposite was observed under CS, where 56% would punish more often.

A hint that collective sanctions somewhat erode morality is suggested by the post-hoc justifications of taking a *blue* envelope, i.e. the 'honest' course of action. In the post-game survey, I suggested three main reasons why people might decide to take the blue envelope, and let them choose an answer. These reasons were, a) the unwillingness to bring harm to a third party ("Taking the green envelope means taking money away from the charity"), b) avoidance of being punished by the external authority ("The green envelope can be empty"), and c) avoidance of being punished by one's peer ("The other participant can make the green envelope empty"). The participants rated the importance of these factors in their decision of taking the blue envelope on a 5-point Likert scale ranging from 1 (less important) to 5 (more important).

Table 3.6 shows the average importance of these motives across treatments. As can be seen, under individual sanctions, the harm to a third party – or the relative morality of the action – was somewhat more important to participants, whereas for the other, more self-interested motives, the importance-rating was similar across sanctions. This is in line with the idea that collective sanctions erode the individual responsibility.

TABLE 3.6: Motives for choosing the *blue* envelope

	Individual Sanctions	**Collective Sanctions**
Taking the green envelope takes money away from the charity	3.8	3.3
The green envelope can be empty	2.8	2.9
The other participant can make the green envelope empty	2.8	2.9

Similarly, I had people rate potential motives why they would *refrain* from punishing their peers. The participants evaluated the relative importance of the following factors: a) individual cost ("Emptying the envelope was too costly"), b) moral considerations ("It is bad to punish other people without sufficient reason"), and c) fear of retaliation ("If I do it, the other participant may empty my envelope as well"). These ratings are shown in Table 3.7.

TABLE 3.7: Motives for *refraining* from punishment

	Individual Sanctions	**Collective Sanctions**
Emptying the envelope was too costly	3.0	3.1
It is bad to punish other people without sufficient reason	3.8	2.9
If I do it, the other participant may empty my envelope as well	2.3	2.9

Again we can see that moral considerations play a more important role under individual sanctions than under collective sanctions.

A final piece of evidence comes from the trust question. I asked respondents the standard trust question ("Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?").

TABLE 3.8: Choice of *blue* envelope depending on self-reported trust

	Individual Sanctions	**Collective Sanctions**
Most people can be trusted.	50.3%	37.2%
Need to be very careful with people.	37.2%	37%

Table 3.8 reports the share of participants that chose the blue envelope/behaved 'honestly' depending on their self-reported trust and the treatment condition. We see that under individual sanctions, there is the positive correlation between trust and 'honest' behavior that we would expect to see: trusting people – who believe in the morality of others – behave more morally themselves. Under collective sanctions, this correlation vanishes. Trusting people behave just as morally as amorally as non-trusting individuals.

3.8 Conclusions

We are now in a position to make a more informed judgment about the effectiveness and efficiency of collective sanctions. It seems clear that if a policy maker would like to curb the crime rate, or somehow differently affect prosocial behavior, it does not seem to be advisable to count on collective sanctions. This is for two reasons.

First, collective sanctions are not particularly effective. The data showed that collective sanctions *did increase* peer punishment and strengthened the link between peer punishment and reduced crime. However, neither in the within-subject design, nor in the between-subject design, did collective sanctions perform better than individual sanctions in actually reducing the crime rate (i.e. increasing donations to charity).

Second, collective sanctions are inefficient. By increasing the rate of punishment, they also caused a lot of money being spent in an unproductive way. As shown, earnings were significantly lower under collective sanctions.

They thus constitute an expensive way to reach a crime rate that can more efficiently be reached through individual sanctions.

Two questions remain open, however. First, why, if collective sanctions are so inefficient, are they still widely used, as shown in the introduction to this paper? Second, what are the exact reasons why collective sanctions underperform despite the fact that they encourage peer punishment? I provided some evidence that the feeling of unfairness generated by collective sanctions undermines moral motivations that also help to drive down crime rates. However, more detailed studies would be necessary to fully answer this question.

Conclusion

The advantage of a correctly conducted experimental approach is that it focuses only on one single property of a phenomenon in question, isolating it from all contextual factors (Guala, 2009). By testing a bare, highly abstract construction against a baseline scenario, we minimize the risk of interactions with other factors. Otherwise, these interferences make it hard to draw causal conclusions (Neuman, 2013, p.282). This strong point of the experimental approach, however, results in a certain degree of artificiality, inevitably leaving many important questions on the sidelines (Webster and Sell, 2014).

In this conclusion, I would like to briefly describe the topics that, due to this study's chosen methods, were not covered enough in this thesis. These topics are a logical continuation of some key questions raised during the course of this thesis' studies.

The discussion of consequences of collective sanctions necessarily revolves around *groups*. Notably, we discussed how collective sanctions affect intergroup cooperation, ingroup cooperation and peer punishment. Except for brief references to feelings of unfairness, questions of *individual* perception and decision-making were largely neglected. This is true in particular for the important question of an individual's decision to enter or leave a group. That makes us ask the following questions: how do collective sanctions affect group composition? Do some conditions make collective sanctions a more legitimate solution in the eyes of their target? Does the group increase its coherence under the threat of collective sanctions or do, in contrast, single members actually feel more atomized? How does the share of sanctions a

given group member receive affect his or her behavior in the group?

These questions of selection, legitimacy, interdependence and distribution are deeply interconnected. Nevertheless, for the sake of clarity, I discuss them one by one.

Selection

Both theoretical and empirical works have shown that selection (both with regard to entry into, and exit from a group) is an important factor in group cooperation; groups with more selective entrance requirements typically demonstrate higher degrees of cooperation (Ahn, Isaac, and Salmon, 2009). In all three of this thesis' studies we treated the groups' composition as given. Individuals received their group membership as an ascribed property in the beginning of the experiment. Although this is the typical design for most of the experimental studies of social dilemmas (Chaudhuri 2011 provides a short list of experiments on endogenous group composition), this static picture is by far not complete, since selection processes can exert a strong influence on the pattern of cooperation established within a group.

An example is provided by Guthrie (2000), who studied employee turnover. He compared companies in which team compensation was equality-based with companies in which team compensation was equity-based. The former can be interpreted as collective reward (with similar properties as collective sanctions), while the latter is closer to traditional rewards based on individual effort. The equality-based groups had a lower turnover than those in which team rewards were distributed based on each member's individual input – the composition in terms of personnel in equity-based teams were much more stable. If this logic can be projected onto different sanctions

regimes, this would suggest that groups operating under individual sanctions should be less sustainable, and those operating under collective sanctions more sustainable. Through this mechanism, collective sanctions may affect not only the longevity of the group, but also its quality. If those who decide to stay or join such a group would tend to have similar properties, they will create a specific culture that in turn will shape who selects into the group.

Legitimacy

Considering such selection dynamics turns our attention to the related question of legitimacy. Narveson (2002), in his treatise on the legitimacy of collective sanctions, classifies groups according to their susceptibility for collective liability. His main classification criterion is the degree of freedom of membership. He argued that in groups where members are free to join and quit at any moment, it makes sense to blame each member for an act committed on behalf of the group by one or some its members. This is because the very act of membership in this case is a demonstration of adherence to group values. Collective sanctions may therefore meet a degree of acceptance and be considered legitimate. In contrast, in groups where membership is mandatory, there is no implicit acceptance of group values, and collective sanctions therefore are seen as illegitimate.[7]

A more general point about considerations of legitimacy in the context of sanctioning is in order. Sociology is a positivist science; it is not concerned with how things *should* be done, but rather with how they *actually happen*,

[7] An important but overlooked question is heterogeneity in terms of the ability to leave a group, and the consequences of this heterogeneity on group composition. What if the ability to exit the group is unevenly distributed among members? Then it may happen that under collective sanctions, the perpetrators leave, and only those members who remain get punished – who in fact are innocent, and should not be targets of such sanctions. Such heterogeneity may therefore further undermine the legitimacy of collective sanctions.

leaving normative questions to the moral philosophers. This approach, as seen in Chapter 1's results, is a bit naive: we found that the perception that a specific sanctioning mechanism is unfair resulted in lower cooperation rates. This result has been confirmed in other studies where procedurally unfair punishment resulted in plummeting cooperation rates (Prooijen, Gallucci, and Toeset, 2008), or where rules set by an egoistic leader did not have the same beneficial effect as rules set by a self-sacrificing leader (Mulder and Nelissen, 2010). Therefore, it is important to see how collective sanctions relate to the question of collective responsibility, for it is entirely possible that sanctions are applied to a group without enough justification that the group is actually responsible for its actions. [8]

In modern ethics, there are two radically opposed points of view with regard to collective sanctions, with a wide range of more moderate opinions in between. On the one side of the spectrum there is the position of Karl Popper and methodological individualists who believe that no form of group responsibility is legitimate. Another, less numerous group of scholars believes that there are situations in which we can hold a group accountable for actions committed by its members. These scholars base their belief on Durkheim's theory of the social act: if there is a social structure – such as a cultural or social norm – that transcends the individual and can shape his/her behavior, and it is created by the dominant culture in the group, then all members of this group are responsible for individual actions. Political theorists such as Hobbes or Rousseau, who have legitimized the subjugation of individual wills under a collective entity, are also cited by the supporters of collective responsibility. According to French (1987), a leading philosopher who wrote extensively on the topic of collective responsibility, there is a class of actions that can only be done by a group (e.g. to 'elect' someone or to 'chip in for

[8] A complicating factor is that in telling that a group is responsible, we implicitly assume that *all group members are responsible equally* – a question that we will consider further in the section on the distribution of responsibility and sanctions.

beer'). Whenever such types of actions result in harmful consequences, each group member who participated in this action is responsible for it, French argues. This logic made Karl Jaspers claim that the Germans are collectively guilty of Nazi (and specifically Hitler's) crimes (Jaspers, 2009). So, we may say that collective sanctions are perceived as unfair only if the mechanism that transfers this guilt to other members is missing. Such mechanisms of transferal are the act of staying within the group, or the delegation of one's will to a wrongdoer.

Often, the survival, safety and well-being of each member depends on others in a group and on that group's ability to act collectively. This brings up another question: Can we ever talk about entirely independent individual actions – or are all actions, due to group members' ever-present interdependence, in fact *group-based* decisions?

Interdependence

Moral philosophers distinguish between aggregate and conglomerate collectives. Aggregate collectives are just collections of people (e.g. a crowd at a bus station waiting for the next bus), while a conglomerate is an "organization of individuals such that its identity is not exhausted by the conjunction of the identities of the persons in the organization" (French, 1987, p.13). According to French, then, the degree of group cohesion is crucial for understanding whether we can or cannot hold a group responsible for its members' actions. Not everybody agrees with this position; some claim that there are situations when disjointed agglomerates of people become collectively responsible due to inaction, as in the classic case of Kitty Genovese (Rosenthal, 2015). We may formulate this particular position in terms of interdependence (May, 2006): if people are deeply dependent on each other in committing individual acts,

then it is logical enough to suppose that they should be held collectively responsible for consequences of these individual acts.

Heckathorn (1988) briefly mentions that the more a group is atomized, the stronger collective sanctions must be in order to be effective. But what if collective sanctions themselves change the degree of group cohesion, making a group more consolidated? Again, there is not much research on attitudes toward group sanctions and interdependence. As a proxy, we can use studies in management which examine the effect of positive rewards for a team. As reported in Haines III and Taggar (2006), attitudes toward team-based rewards (as compared to individual effort-based rewards) are more positive in groups with higher degrees of task-interdependence. Other research shows that when a team is collectively rewarded, that increases the amount of helping behavior between members (Bamberger and Levi, 2009). These findings suggest that the form of reward – and, perhaps, the form of sanction – can directly impact group cohesion. However, yet another team of scholars added a further aspect to this story: they showed that collective rewards increased the tendency of teams to overstate performance reports (Maas and Van Rinsum, 2013). This latter finding is an important hint that collective rewards and sanctions can also affect the frequency of norm violations and whistleblowing in groups.

Distribution

We hold it as given that collective sanctions, if they are presented as an institutional mechanism, are applied to all members of the group uniformly, or at least with uniform probability. I began Chapter 2 with the example of Frank Roque who, in his desperation to retaliate against Muslims, killed an innocent Sikh. It would, however, be an oversimplification to assume most people perceive such a degree of outgroup entitativity that they are indifferent to the

death of Bin Laden versus a random Muslim. This brings up the question of the distribution of sanctions: depending on different distributional schemes, people may consider outgroup members more or less responsible for the actions committed by other members of the outgroup. There is a consensus among moral philosophers that individuals cannot be responsible for group actions if they openly dissent with these decisions (French, 1987). If we talk about groups of substantial size, it is logical to assume that a norm violator's nearest social neighborhood is considered to hold more responsibility than outgroup members who are not even acquainted with him. For example, we may imagine a system where an individual's closest friends tend to be punished for that individual's misdeeds, which may cause that individual to lose his social links. Depending on the punisher, this may be the desired outcome.

Besides social or physical distance, there are also other criteria to decide who in a group should be punished for that group's misdeed. For example, one common behavior is that an individual is selected for punishment who is actually *able to pay the punishment's price*. Feinberg (1968), in his treatise on collective responsibility, describes a case similar to the case of Kitty Genovese. He invites the reader to imagine a situation of a man swimming off a public beach that lacks a professional life-saver. That man 'shouts for help in a voice audible to a group of one thousand accomplished swimmers lolling on the beach; and yet no one moves to help him, and he is left to drown.' If the widow of the drowned man has to choose one swimmer to sue for negligence, Feinberg wonders, whom would she pick? The most logical decision, from the point of view of rational outcomes, would be to choose the richest one. This logic, joined with the ability of people to leave a group, can produce unexpected consequences.

A significant part of microsociology and behavioral economics is dedicated to the study of social dilemmas and the study of ways of reducing the tension between individual and collective rationality (Kollock, 1998). Since

late 1950s, there have been hundreds, if not thousands, of studies that use game-theoretical models and experiments based on these models to examine how human groups are able to overcome the egoistic desires of individual members. Some of the mechanisms identified in this research program are rewards, punishment, selection and the selective exclusion of group members (Sally, 1995; Balliet, Mulder, and Van Lange, 2011; Sasaki and Uchida, 2013).

At least for groups that do not pursue criminal goals, higher cooperation levels achieved within a group generally result in positive changes for society as a whole.[9] Thus, the prosperity of individual members is tightly connected to the prosperity of the group. The opposite is also true: the well-being of individual members adds up to the wellbeing of the group. That is why most behavioral studies are focused on finding institutional designs that are effective in solving social dilemmas. However, our inclination to find a way to increase group cohesion is not congruous with the impulsive denial of collective responsibility that is also typical for many social scientists. If the ability to act collectively increases the chances that individual members of a group violate external norms with impunity, then the entire group is partially responsible for such actions. Similarly, if the cooperative environment within a group is directly associated with hostility towards outsiders, then, yet again, each group member who made his contribution to the creation of such an environment is somewhat responsible for the adverse actions of his or her peer.

One does not need to be a social scientist to instinctively understand that we all owe our existence to the multitudinous groups we belong to, both because they help us to construct our social identity (Turner and Oakes, 1986) and because as a biological species, it is just easier to survive within a large

[9]For example, solidarity is thought to be instrumental in solving issues such as depletion of scarce resources (Ostrom, 2015).

cooperative group (Kokko, Johnstone, and Clutton-Brock, 2001). But this folk wisdom, and reckless adherence to the maxim that 'no man is an island' can be drivers of dangerous prejudice and stereotyping against members of other groups. To prevent this, it is crucial to continue the studies of collective responsibility and sanctions and to develop a deeper understanding of their consequences.

Appendix A

Experimental instructions

After the experiment, your total earnings from the experiment will be paid out to you anonymously and in cash.

The following pages describe in detail the experiment.

The experiment is divided into different periods. There will be 15 periods in total. During all 15 periods, the participants are divided into groups of three. Therefore, you will be in a group with 2 other participants. The composition of the groups will remain the same during all the experiment.

Each period consists of three stages. In the first stage, you have to decide how many tokens you contribute to a group project. In the second stage, there is a chance that the contributions to the group project by all group members are checked by the computer. If they are, everyones earnings are reduced if at least one group members contribution is below a specific amount. In the third stage, you will learn how much the other members of your group contributed to the project and decide whether to reduce or leave equal the earnings of each other group members.

A.1 The first stage

At the beginning of each period, each participant in your group receives 20 tokens. We will refer to these tokens as the endowment.

In the first stage, you decide how to use your endowment. You have to choose how many tokens you want to contribute to a group project and how many of them to keep for yourself. You can contribute any amount of your endowment to the group project. How many tokens you contribute is up to you. Each other group member will also make such a decision. All decisions are made simultaneously. That is, nobody will be informed about the decision of

You are now taking part in an economic experiment. You can earn money depending on your decisions and the decisions of other participants. How you can earn money is described in these instructions. It is therefore important that you read these instructions carefully.

During the experiment, you are not allowed to communicate with other participants in any way. If you have any questions please raise your hand. One of us will come to your table to answer your question. During the experiment, your earnings will be calculated in tokens. At the end of the experiment, the total amount of tokens you have earned will be converted to US dollars at the following rate:

$$10 tokens = 50 cents$$

the other group members before everyone has made his or her decision.

Your earnings in tokens, in each period, are the sum of two parts:

1. The number of tokens that you have kept for yourself.

2. Your income from the group project. This income is calculated as follows:

Income from the group project = 0.5 sum of contributions of all group members to the project Notice that, for each token that you keep for yourself you earn 1 token. If instead you contribute this token to the group project,

then the total contribution to the project will go up by one token. Your income from the group project will go up by 0.5 tokens. Moreover, the other group members? income from the project will also go up by 0.5 tokens. Your contribution to the group project therefore also increases the income of the other group members. For each token contributed to the project, the total earnings of the group will rise by 1.5 tokens. Note that, you also earn tokens for each token contributed to the group project by the other group members. For each token contributed by any member, you earn 0.5 tokens.

In summary, your earnings in tokens at the first stage of a period are equal to:

20 your contribution + 0.5 (sum of contributions)

After everyone has made his or her decision the first stage ends.

A.2 Example for the first stage

Here is an example that illustrates how the earnings in tokens are calculated in the first stage of each period. The numbers used in the example were chosen arbitrarily.

You are in a group with two other participants (group member 1 and group member 2). You contribute 15 tokens to the group project, group member 1 contributes 5 tokens, and group member 2 contributes 10 tokens.

- In this case, your earnings equal: $20\ 15 + 0.5(15 + 5 + 10) = 20$ tokens.
- Group member 1s earnings equal: $20\ 5 + 0.5(15 + 5 + 10) = 30$ tokens.
- Group member 2s earnings equal: $20\ 10 + 0.50(15 + 5 + 10) = 25$ tokens. The second stage In the second stage, there is a 33% chance that the contributions of everyone in your group are checked by a computer.

Specifically, every period the computer generates a random number between 1 and 100 for each group. If the generated number equals 33 or less then it checks the contributions of all group members in that group.

- If your groups contributions are checked and the contribution of at least one group member is found to be 10 tokens or less, then the earnings of everyone in the group in that period are reduced by 7 tokens.
- If your groups contributions are checked and the contribution of each of the members in your group is found to be 11 tokens or more, then the earnings of everyone in the group remain the same.
- If your groups contributions are not checked then the earnings of everyone in the group remain the same. Example for the second stage As in the previous example, you are in a group with two other participants (group member 1 and group member 2). In the first stage, you contribute 15 tokens to the group project, group member 1 contributes 5 tokens, and group member 2 contributes 10 tokens. At the end of the first stage the earnings of your group members are as follows:
  - Your earnings are: 20 tokens
  - Group member 1s earnings: 30 tokens
  - Group member 2s earnings: 25 tokens.

If the generated number in this period is 25 then the computer checks the contributions and the earnings for the second stage equal:

- Your earnings are: 20 7 = 13 tokens
- Group member 1s earnings: 30 7 = 23 tokens
- Group member 2s earnings: 25 7 = 18 tokens. The third stage At the beginning of the third stage, everyone in the group will see how much

each of the other group members contributed to the project, whether the group was checked by a computer or not, and their earnings at the end of the second stage. The decision each group member has to make in the third stage is to either reduce or leave equal the earnings of each other group member. The other group members can also reduce your earnings if they wish to. All decisions are made simultaneously. That is, nobody will be informed about the decision of the other group members before everyone made his or her decision. More concisely, in this stage you must decide how many deduction points you want to allocate to each of the other two group members. For each deduction point that you allocate to another group member, his or her earnings are reduced by 2 tokens and your own earnings are reduced by 1 token. If you do not wish to change the earnings of another group member then you must allocate 0 deduction 21 points to him or her. Each participant can allocate up to 10 deduction points on each group member in each period. After everyone has made a decision, you will be informed how many deduction points the other group members allocated to you and what your total earnings for that period are. Note that you will only be informed of the total amount of deduction points allocated by the other two group members. You will not know how many deduction points each individual group member allocated to you.

In summary, your earnings in tokens at the third stage of a period are equal to:

Second stage earnings ?2× deduction points others allocated to you ? deduction points you allocated

A.3 Example for the third stage

Here is an example that illustrates how your earnings are calculated in the third stage.

You are in a group with two other participants (group member 1 and group member 2). Suppose that after the second stage you have earnings that are equal to 30 tokens. In the second stage you decide to allocate 3 deduction points to group member 1 (this reduces group member 1?s earnings by 6 tokens) and 0 deduction points to group member 2 (this does not change group member 2?s earnings). After all have made their decision, you learn that the others allocated you a total of 4 deduction points. Your total earnings in tokens in this period are then equal to: $30?2 \times 4?3 = 19$ tokens.

Appendix B

Screenshots of an Experiment in CELSS

(see separate pages below)

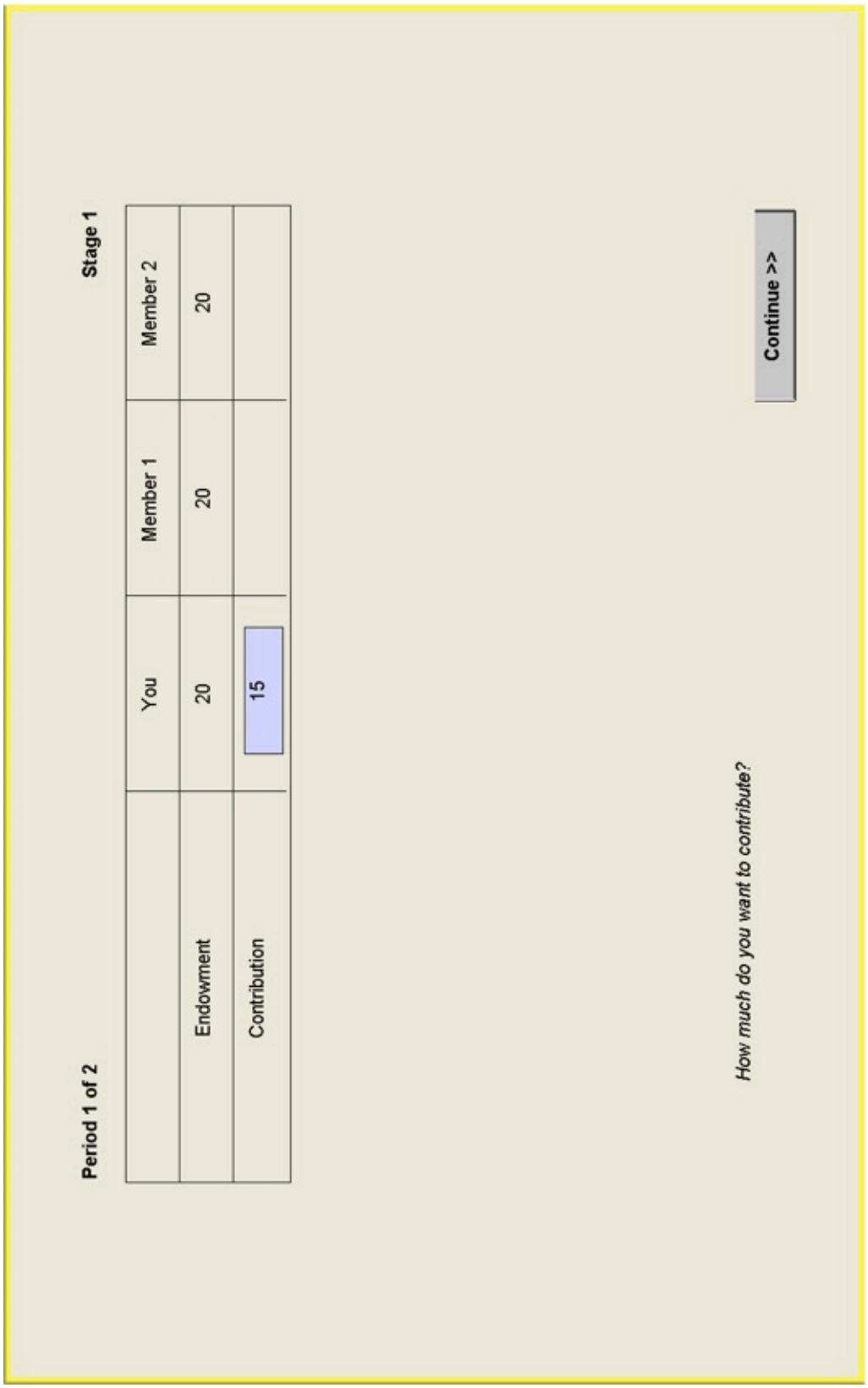


FIGURE B.1: Screenshot1

FIGURE B.2: Screenshot2

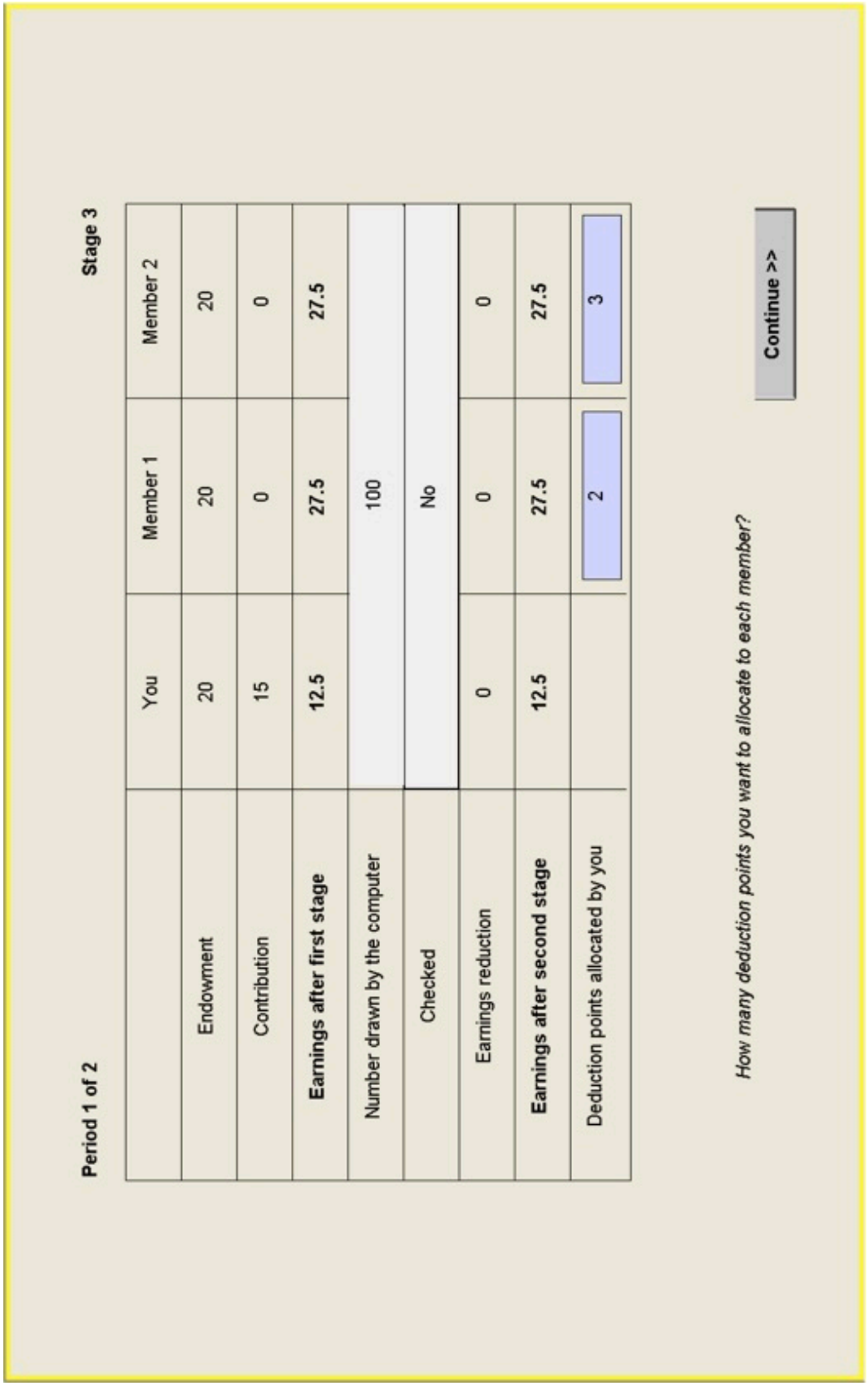


FIGURE B.3: Screenshot3

Appendix C

Lab Instructions

Welcome and thank you for participating in this experiment. Please take your time to read through the following instructions thoroughly. You can take notes if you so wish.

In this study you will earn some money. As a participant you will be asked to make certain decisions, and to complete a brief questionnaire. There are no right or wrong decisions in this experiment. You will each receive a show-up fee of 100 Rs, and you will earn an additional amount of money. This could be between 100 Rs and 300 Rs, depending on the decisions that you and others make. You will be paid at the end of the workshop.

Participation in this study is voluntary. After the details of the study have been explained to you, you may decline to participate if you so wish. However please note that if you choose not participate, you will receive only your show-up fee.

All the decisions you make and information you provide will be treated as confidential. Your name will not be included or in any other way associated with the data collected in the study. All the monetary amounts mentioned below are Indian rupees. At the end of the experiment your earnings will be paid in cash. Instructions for 'Collective Punishment' treatment (p=10%)

In these experiments the participants earn real money. At the end of each round you can see at the screen how much you earned – in rupees.

At the end of this session you will receive the show-up fee of 100 rupees plus the amount you earned during the experiment. We will also ask you to fill in a short questionnaire. We will ask you to provide your name and surname in order to know what amount to pay you at the end.

The experiment consists of twenty small rounds. Each round lasts from 30 seconds to one minute depending on how fast you and other participants will make the decisions. In total the experiment will not take more than one hour of your time including the questionnaire and payment.

In each round you will be a part of a group of **two**. In every round you will be with a **different** person (your co-participant will change).

You will not know with whom you are in the group each round. The study is completely anonymous. In every round you will see the screen with 20 envelopes of two different colors: 10 blue and 10 green. [The Picture 1 is shown – see below]

You should choose one of the envelopes by clicking on the one you have chosen and then pressing OK. Unless you have not clicked 'OK' button you can change your mind by clicking another envelope of your choice. We do not limit you in decision making - please take your time.

In each **blue** envelope there are 10 rupees.

In 9 out of 10 **green** envelopes there are 25 rupees.

In 1 out of 10 **green** envelopes there are 0 rupees[1].

These empty envelopes are random each round.

We have a certain amount of money that we will send to a charity organization in India after the experiment[2]. Each time someone of you chooses the

[1] In a treatment with p=30% these two sentences were changed to:
"In 7 out of 10 green envelopes there are 25 rupees.
In 3 out of 10 green envelopes there are 0 rupees."

[2] Later on participants were informed about the details of a charity organization. That was "Community Service Initiative – Prayyas", supported by Indian Institute of Management – Ahmedabad. See details at: http://www.iimahd.ernet.in/institute/campus/student-activities/clubs/prayaas.html

green envelope the amount we send to the charity organization will diminish by 15 rupees.

Each time you chooses the **green** envelope and it turns out to be **empty** the earnings of your partner will diminish by 10 rupees[3].

So each time your partner chooses the **green** envelope and it turns out to be empty, your earnings will diminish by 10 rupees.

You and your partner-participant have made your decisions on which envelope to choose. Then at the next screen both of you can “catch" the other participant for his/her potential choice of the **green** envelope making it “empty". But it will cost you 5 rupees (the second participant can make the same with you).

If the other participant chose the **empty green** envelope and you choose the option “Make the green envelope of your partner empty", your earnings will not diminish by 10 rupees (but the “emptying" will cost you 5 rupees).

If you choose the option “Make the green envelope of your partner empty" and the other participant chose the **non-empty green** envelope, he will get 0 rupees (but the “emptying" will cost you 5 rupees).

If the other participant chose the **blue** envelope and you choose the option “Make the green envelope of your partner empty", the “emptying" will not cost you 5 rupees and your partner receives what is due to him.

[3]In Individual Punishment treatment this and the next sentences were omitted.

FIGURE C.1: Screenshot

Appendix D

Sessions

TABLE D.1: Table D1

Date	Time	N	Treatment	P
Jan 29	5pm	18	IP-CP	0.1
Jan 29	7pm	18	IP-CP	0.1
Feb 2	5pm	14	CP-IP	0.1
Feb 2	7pm	8	CP-IP	0.1
Feb 3	6pm	14	CP-IP	0.3
Feb 3	7:30pm	8	IP-CP	0.3
Feb 10	4pm	16	CP	0.3
Feb 10	7pm	12	IP	0.3
Feb 12	4pm	8	CP	0.3
Feb 12	6pm	20	IP	0.3
Feb 12	7pm	12	CP	0.3
Feb 13	4pm	12	IP	0.3

Appendix E

Answers to some post-experimental questions

If there would be LESS empty green envelopes, would you take them MORE often? (%)

TABLE E.1: Table E1

	Individual Sanctions	Collective Sanctions	**Total**
Yes	50	58.3	**53.7**
No	50.1	41.6	**46.2**

If there would be LESS empty green envelopes, would you make other people's envelope empty LESS often?

TABLE E.2: Table E2

	Individual Sanctions	Collective Sanctions	**Total**
No	65.9	63.9	**65.0**
Yes	34.1	36.1	**35.0**

In the cases when you decided NOT to take green envelope, what were the reasons for it (1– less important, 5 – more important)? (Standard errors in parentheses)

TABLE E.3: Table E3

	Individual Sanctions	**Collective Sanctions**
Taking green envelope takes	3.7	3.3
money from charity	(.2)	(0.25)
Green envelope can be empty	2.7	2.9
	(.21)	(.23)
Another participant can	2.6	2.8
make envelope empty	(.21)	(.24)

In the cases when you decided NOT to EMPTY an envelope of another participant, what were the reasons for it (from more to less important)?

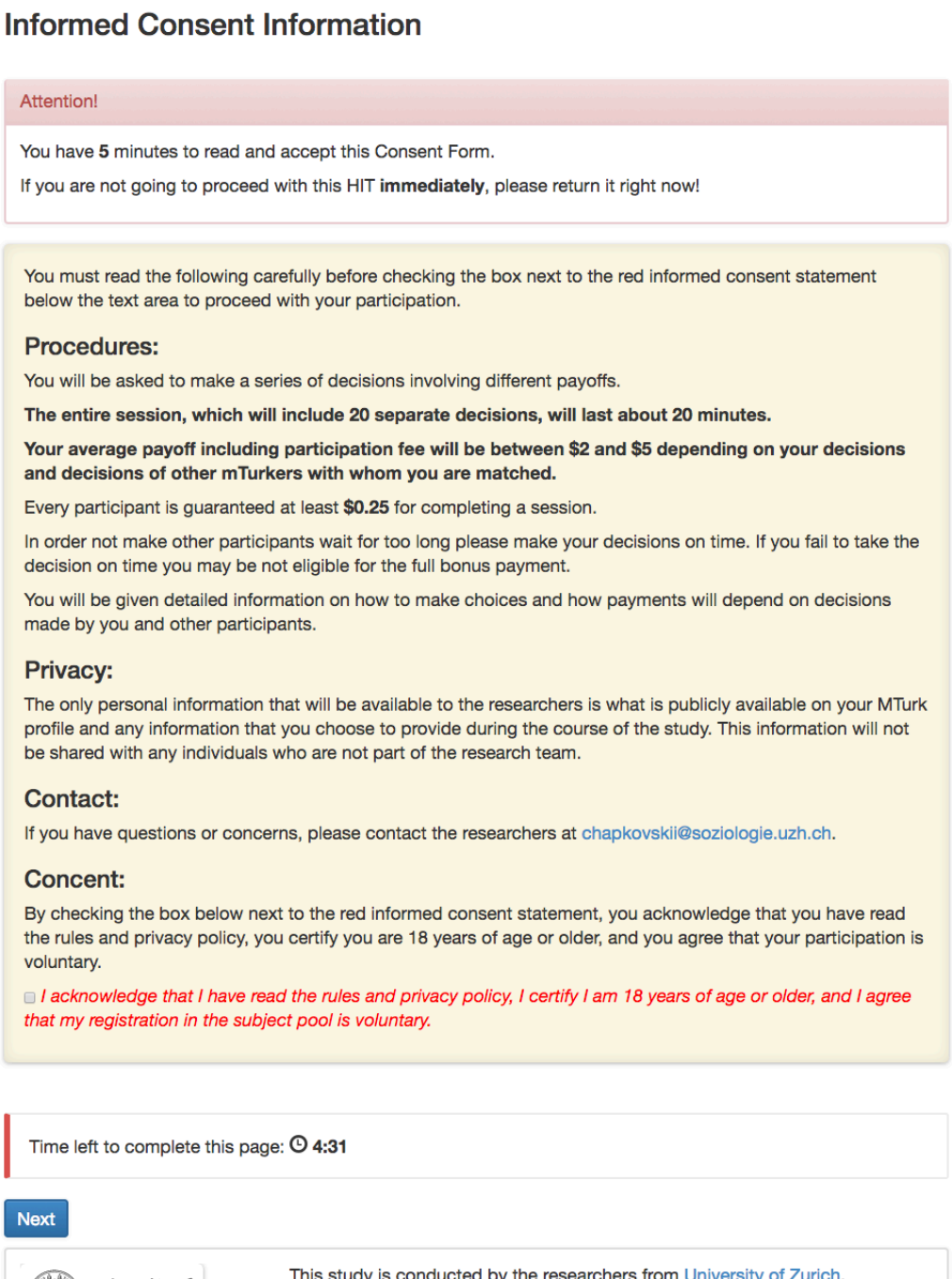
TABLE E.4: Table E4

	Individual Sanctions	**Collective Sanctions**
Emptying envelope was too costly	2.9	3.1
	(.21)	(0.23)
It is bad to punish other people	3.7	2.8
without sufficient reason	(.19)	(.26)
If I do it someone else can	2.3	2.9
'empty' my envelope as well	(.21)	(.28)

Appendix F

Chapter2 Appendix

F.1 Consent Form



Informed Consent Information

Attention!

You have **5** minutes to read and accept this Consent Form.

If you are not going to proceed with this HIT **immediately**, please return it right now!

You must read the following carefully before checking the box next to the red informed consent statement below the text area to proceed with your participation.

Procedures:

You will be asked to make a series of decisions involving different payoffs.

The entire session, which will include 20 separate decisions, will last about 20 minutes.

Your average payoff including participation fee will be between $2 and $5 depending on your decisions and decisions of other mTurkers with whom you are matched.

Every participant is guaranteed at least **$0.25** for completing a session.

In order not make other participants wait for too long please make your decisions on time. If you fail to take the decision on time you may be not eligible for the full bonus payment.

You will be given detailed information on how to make choices and how payments will depend on decisions made by you and other participants.

Privacy:

The only personal information that will be available to the researchers is what is publicly available on your MTurk profile and any information that you choose to provide during the course of the study. This information will not be shared with any individuals who are not part of the research team.

Contact:

If you have questions or concerns, please contact the researchers at chapkovskii@soziologie.uzh.ch.

Concent:

By checking the box below next to the red informed consent statement, you acknowledge that you have read the rules and privacy policy, you certify you are 18 years of age or older, and you agree that your participation is voluntary.

☐ *I acknowledge that I have read the rules and privacy policy, I certify I am 18 years of age or older, and I agree that my registration in the subject pool is voluntary.*

Time left to complete this page: 4:31

Next

University of

This study is conducted by the researchers from University of Zurich.

F.2 Waiting Page

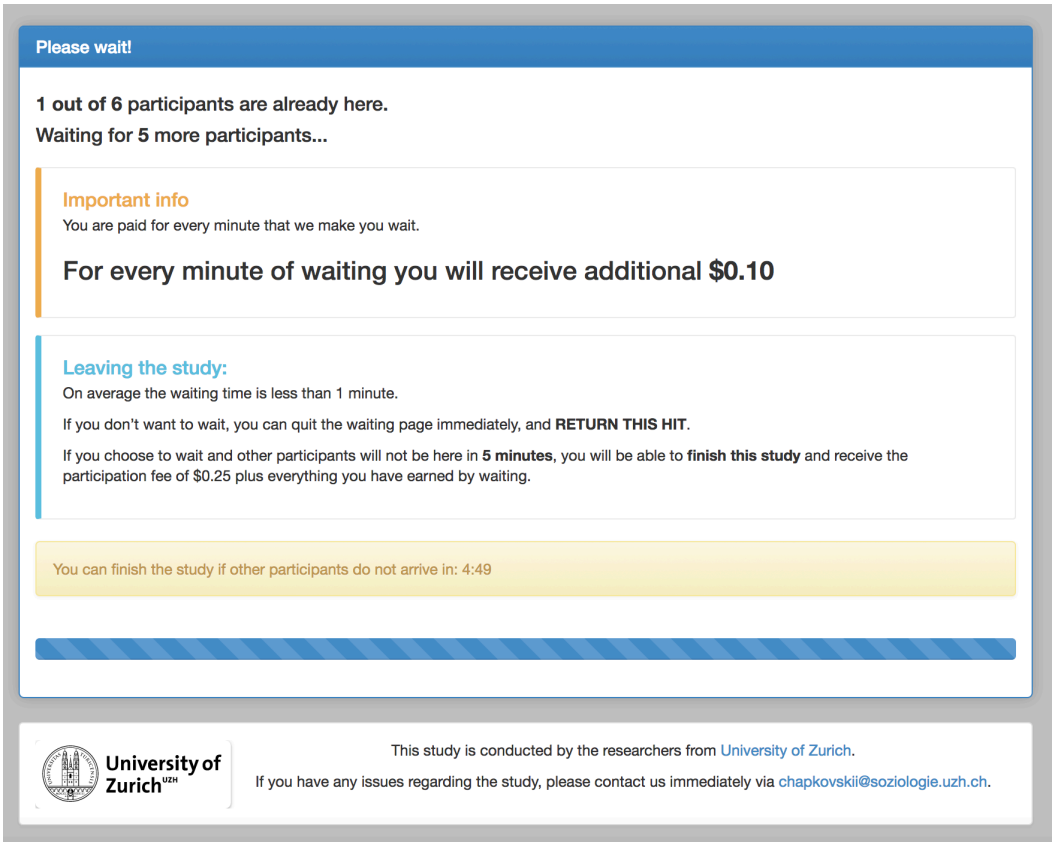


FIGURE F.2: Waiting Page

F.3 Stage 1 Decision

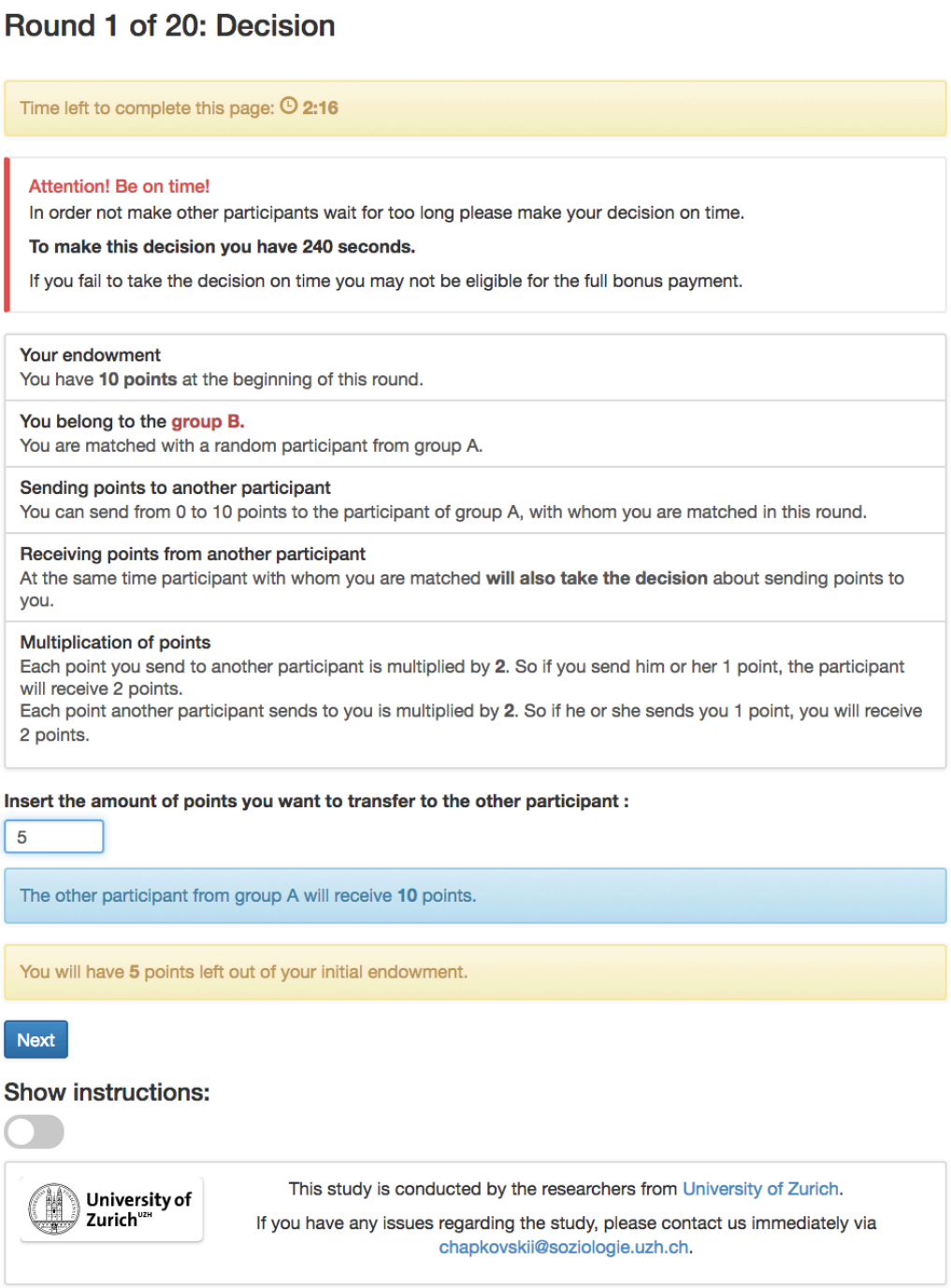


FIGURE F.3: Stage 1 Decision

F.4 Stage 2 Decision

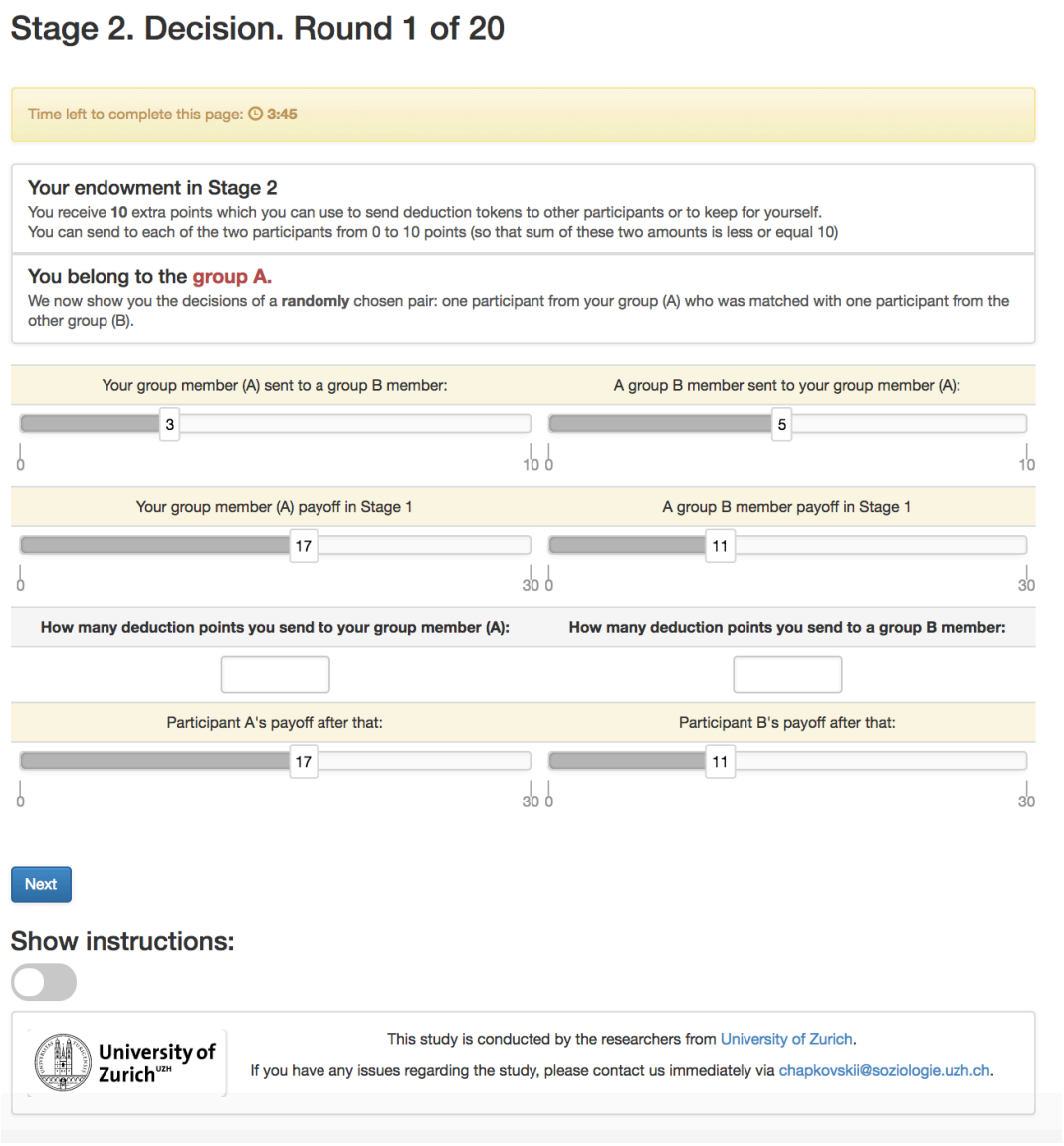


FIGURE F.4: Stage 2 Decision

F.5 Stage 1 Instructions

Time left to complete this page: **4:48**

There are two stages in this study:

General Rules

In this game there are six participants. Each participant is randomly assigned to a group of three. There are two groups: A and B, with three participants in each group.

In each round every participant from the group A is matched with a random participant from the group B. They then play a game, in which both of them should take a decision that is described below.

You have been assigned to the **group B.**

In each round you are matched with one **random** participant from **group A**.

The experiment consists of **20** identical rounds.

Stage 1:

In each round you are given **10 points** and you have an option to transfer some of these points or all of them to the other participant, with whom you are matched, or leave all the points to yourself. The amount of points you decide to transfer to the other participant will be multiplied by **2** and earned by the other participant. Simultaneously the other participant will be given exactly the same choice.

For instance:

- If you both transfer all the points you get, each of you will get 20 points.
- If you transfer everything and the other participant keeps everything, then you will earn nothing and the other participant will earn 30.
- If you keep everything and the other participant transfers everything, then you will get 30 points and your partner will earn nothing.

Payoffs

In the end of the experiment your earnings will be calculate as follows:

You will get a participation fee of $0.25.

In addition all your payoffs in all 20 rounds will be summed up and converted in US dollars at the exchange rate of:

200 points = 1 US dollar.

FIGURE F.5: Stage 1 Instructions

F.6 Stage 2 Instructions

In each round in Stage 2 you again have an endowment of 10 points.

In this stage you can see the decisions and thus payoffs of another pair of participants in the same game, i.e. of a random member of your own group and a member of the other group, with whom the member of your group was matched. After seeing the results of their game, you can transfer points from your endowment to decrease the income of the participants. By transferring 1 point from your endowment you decrease the income of another participant by 3 points. **While in your own group your action decreases the payoff of the specific participant, whose decisions you have just seen, in the other group your action will affect a random participant of the other group.**

You can not send more than 10 points to both participants in total.

Other members of both groups can make the same decisions about other participants. When you make your decision in Stage 2 and click 'Next', you will be shown your own and your partner's decisions made in Stage 1 and your payoffs for this round.

These instructions remain available to you at the later stages of the experiment.

FIGURE F.6: Stage 2 Instructions

Bibliography

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner (2002). "An experimental bribery game". In: *Journal of Law, Economics, and Organization* 18.2, pp. 428–454. URL: http://jleo.oxfordjournals.org/content/18/2/428.short (visited on 03/02/2016).

Abeler, Johannes, Anke Becker, and Armin Falk (2014). "Representative evidence on lying costs". In: *Journal of Public Economics* 113.Supplement C, pp. 96–104. ISSN: 0047-2727. DOI: 10.1016/j.jpubeco.2014.01.005. URL: http://www.sciencedirect.com/science/article/pii/S0047272714000061 (visited on 01/06/2018).

Agrawal, Arun and Sanjeev Goyal (2001). "Group Size and Collective Action: Third-party Monitoring in Common-pool Resources". en. In: *Comparative Political Studies* 34.1, pp. 63–93. ISSN: 0010-4140. DOI: 10.1177/0010414001034001003. URL: https://doi.org/10.1177/0010414001034001003 (visited on 01/06/2018).

Ahn, Toh-Kyeong, R. Mark Isaac, and Timothy C. Salmon (2009). "Coming and going: Experiments on endogenous group sizes for excludable public goods". In: *Journal of Public Economics* 93.1-2, pp. 336–351.

Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew (1954). "The nature of prejudice". In:

Anderson, Christopher M. and Louis Putterman (2006). "Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism". In: *Games and Economic Behavior* 54.1, pp. 1–24. ISSN: 0899-8256. DOI: 10.1016/j.geb.2004.08.007. URL: http:

//www.sciencedirect.com/science/article/pii/S0899825604001654 (visited on 03/09/2016).

Andreoni, James (1988). "Why free ride?: Strategies and learning in public goods experiments". In: *Journal of public Economics* 37.3, pp. 291–304.

Andreoni, James and Ragan Petrie (2008). "Beauty, gender and stereotypes: Evidence from laboratory experiments". In: *Journal of Economic Psychology* 29.1, pp. 73–93.

Baldassarri, Delia and Guy Grossman (2011). "Centralized sanctioning and legitimate authority promote cooperation in humans". en. In: *Proceedings of the National Academy of Sciences* 108.27, pp. 11023–11027. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1105456108. URL: http://www.pnas.org/content/108/27/11023 (visited on 01/08/2018).

Balliet, Daniel, Laetitia B. Mulder, and Paul AM Van Lange (2011). *Reward, punishment, and cooperation: a meta-analysis.* American Psychological Association.

Balliet, Daniel and Paul AM Van Lange (2013). "Trust, conflict, and cooperation: A meta-analysis." In: *Psychological Bulletin* 139.5, p. 1090.

Balliet, Daniel et al. (2011). *Sex differences in cooperation: a meta-analytic review of social dilemmas.* American Psychological Association.

Bamberger, Peter A. and Racheli Levi (2009). "Team-based reward allocation structures and the helping behaviors of outcome-interdependent team members". In: *Journal of Managerial Psychology* 24.4, pp. 300–327. URL: http://www.emeraldinsight.com/doi/abs/10.1108/02683940910952705 (visited on 03/02/2016).

Barnes, Steven A. (2011). *Death and Redemption: the Gulag and the Shaping of Soviet society*. Princeton University Press. URL: https://books.google.com/books?hl=en&lr=&id=kewLQwngUSkC&oi=fnd&pg=PP2&dq=4.%09Barnes,+Steven+A.+2011.+Death+and+Redemption:+the+Gulag+and+the+

Shaping+of+Soviet+society.+Princeton+University+Press.&ots=1v-LUAAlFB&sig=oGjtaM_a_CLI7FR5vdKrsV5Dn9U (visited on 03/02/2016).

Becker, Gary S. and Richard A. Posner (2009). *Uncommon sense: economic insights, from marriage to terrorism*. University of Chicago Press.

Bellemare, Charles, Luc Bissonnette, and Sabine Kröger (2014). "Statistical power of within and between-subjects designs in economic experiments". In: *Cahier de recherche/Working Paper* 14, p. 25. URL: https://depot.erudit.org/bitstream/004001dd/1/CIRPEE14-25.pdf.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012). "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk". In: *Political Analysis* 20.3, pp. 351–368. ISSN: 1047-1987. URL: http://www.jstor.org/stable/23260322 (visited on 12/31/2017).

Bernhard, Helen, Urs Fischbacher, and Ernst Fehr (2006). "Parochial altruism in humans". In: *Nature* 442.7105, p. 912. URL: http://search.proquest.com/openview/72792168e7b8d487deebe46cbd4c9e73/1?pq-origsite=gscholar&cbl=40569 (visited on 10/04/2017).

"Bihar's law" (2016). "Bihar's draconian prohibition law". In: *The Hindu*. URL: http://www.thehindu.com/opinion/editorial/Bihar%E2%80%99s-draconian-prohibition-law/article14551111.ece.

Bochet, Olivier, Talbot Page, and Louis Putterman (2006). "Communication and punishment in voluntary contribution experiments". In: *Journal of Economic Behavior & Organization* 60.1, pp. 11–26. ISSN: 0167-2681. DOI: 10.1016/j.jebo.2003.06.006. URL: http://www.sciencedirect.com/science/article/pii/S0167268105001101 (visited on 01/05/2018).

Brandts, J. and G. Charness (2011). "The strategy versus the direct-response method: a first survey of experimental comparisons". In: *Experimental Economics* 14.3, pp. 375–398.

Brewer, Marilynn B (1999). "The psychology of prejudice: Ingroup love and outgroup hate?" In: *Journal of social issues* 55.3, pp. 429–444.

Broadhead, Robert S et al. (2002). "Increasing drug users' adherence to HIV treatment: results of a peer-driven intervention feasibility study". In: *Social Science & Medicine* 55.2, pp. 235–246.

Brockhaus, F.I. and I. A. Efron (1896). "Collegiate Dictionary Vol. XVIIIA (36)". In: *St. Petersburg.: Semyonovskaya tipolithographic (IA Efron)*.

Cadsby, C. Bram and Elizabeth Maynes (1998). "Gender and free riding in a threshold public goods game: Experimental evidence". In: *Journal of Economic Behavior & Organization* 34.4, pp. 603–620. ISSN: 0167-2681. DOI: 10.1016/S0167-2681(97)00010-3. URL: http://www.sciencedirect.com/science/article/pii/S0167268197000103 (visited on 03/02/2016).

Campbell, Donald T. (1958). "Common fate, similarity, and other indices of the status of aggregates of persons as social entities". In: *Systems research and behavioral science* 3.1, pp. 14–25. URL: http://onlinelibrary.wiley.com/doi/10.1002/bs.3830030103/full (visited on 10/09/2017).

Capraro, Valerio, Jillian J. Jordan, and David G. Rand (2014). "Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments". In: *Scientific Reports* 4. ISSN: 2045-2322. DOI: 10.1038/srep06790. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4210943/ (visited on 09/24/2017).

Carpenter, Jeffrey P. (2007a). "Punishing free-riders: How group size affects mutual monitoring and the provision of public goods". In: *Games and Economic Behavior* 60.1, pp. 31–51. URL: http://www.sciencedirect.com/science/article/pii/S089982560600145X (visited on 03/02/2016).

— (2007b). "The demand for punishment". In: *Journal of Economic Behavior & Organization* 62.4, pp. 522–542. URL: http://www.sciencedirect.com/science/article/pii/S0167268106000175 (visited on 03/02/2016).

Casari, Marco (2005). "On the design of peer punishment experiments". In: *Experimental Economics* 8.2, pp. 107–115.

Casari, Marco and Luigi Luini (2012). "Peer punishment in teams: expressive or instrumental choice?" In: *Experimental Economics* 15.2, pp. 241–259. URL: http://www.springerlink.com/index/32q7m16438224021.pdf.

Cason, Timothy N. and Feisal U. Khan (1999). "A laboratory study of voluntary public goods provision with imperfect monitoring and communication". In: *Journal of development Economics* 58.2, pp. 533–552.

Chaudhuri, Ananish (2011). "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature". In: *Experimental Economics* 14.1, pp. 47–83.

Choi, Jung-Kyoo and Samuel Bowles (2007). "The Coevolution of Parochial Altruism and War". en. In: *Science* 318.5850, pp. 636–640. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1144237. URL: http://science.sciencemag.org/content/318/5850/636 (visited on 10/04/2017).

Cohn, Alain, Michel André Maréchal, and Thomas Noll (2015). "Bad boys: How criminal identity salience affects rule violation". In: *The Review of Economic Studies* 82.4, pp. 1289–1308.

Coleman, James Samuel (2000). *Foundations of Social Theory*. Belknap Press of Harvard Univ. Press.

Corlett, J. Angelo (1992). "Collective punishment and public policy". In: *Journal of Business Ethics* 11.3, pp. 207–216. URL: http://link.springer.com/article/10.1007/BF00871968 (visited on 12/25/2016).

Culnan, Mary J. and Pamela K. Armstrong (1999). "Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation". In: *Organization science* 10.1, pp. 104–115. URL: http://pubsonline.informs.org/doi/abs/10.1287/orsc.10.1.104.

Cushman, Fiery, A. J. Durwin, and Chaz Lively (2012). "Revenge without responsibility? Judgments about collective punishment in baseball". In: *Journal of Experimental Social Psychology* 48.5, pp. 1106–1110. URL: http:

//www.sciencedirect.com/science/article/pii/S0022103112000595 (visited on 03/02/2016).

Dickson, Eric S. (2007). *On the (in) effectiveness of collective punishment: An experimental investigation*. Tech. rep. Working paper, New York University. URL: http://www.nyu.edu/gsas/dept/politics/faculty/dickson/dickson_collectivepunishment.pdf (visited on 03/02/2016).

Dollard, John et al. (1939). "Frustration and aggression." In:

Eagly, Alice H. and Wendy Wood (1999). "The origins of sex differences in human behavior: Evolved dispositions versus social roles." In: *American psychologist* 54.6, p. 408.

— (2011). "Social role theory". In: *Handbook of theories in social psychology* 2, pp. 458–476.

Ellingworth, James (2014). *Sanctions start to have effect – on racism in Russian football*. URL: http://rbth.com/sport/2014/10/16/sanctions_start_to_have_effect_on_racism_in_russian_football_40669.html (visited on 12/27/2016).

Erev, Ido, Gary Bornstein, and Rachely Galili (1993). "Constructive intergroup competition as a solution to the free rider problem: A field experiment". In: *Journal of Experimental Social Psychology* 29.6, pp. 463–478.

Falk, Armin, Ernst Fehr, and Urs Fischbacher (2005). "Driving forces behind informal sanctions". In: *Econometrica* 73.6, pp. 2017–2030. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2005.00644.x/full.

Falk, Armin and Urs Fischbacher (2002). ""Crime" in the lab-detecting social interaction". In: *European Economic Review* 46.4, pp. 859–869. URL: http://www.sciencedirect.com/science/article/pii/S0014292101002203 (visited on 12/09/2016).

Fatas, Enrique and Guillermo Mateu (2015). "Antisocial punishment in two social dilemmas". In: *Frontiers in Behavioral Neuroscience* 9. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2015.00107. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4413789/ (visited on 01/06/2018).

Fatas, Enrique, Antonio J. Morales, and Paloma Ubeda (2010). "Blind Justice: An experimental analysis of random punishment in team production". In: *Journal of Economic Psychology* 31.3, pp. 358–373. URL: http://www.sciencedirect.com/science/article/pii/S0167487010000127 (visited on 03/02/2016).

Fearon, James D. and David D. Laitin (1996). "Explaining interethnic cooperation". In: *American political science review* 90.4, pp. 715–735. URL: https://www.cambridge.org/core/journals/american-political-science-review/article/explaining-interethnic-cooperation/CE9BC6184CEB72ECD6E18E17041BAB

Fehr, Ernst and Urs Fischbacher (2004). "Social norms and human cooperation". In: *Trends in Cognitive Science* 8.4, pp. 185–190.

Fehr, Ernst and Simon Gächter (1999). "Cooperation and punishment in public goods experiments". In: *Institute for Empirical Research in Economics working paper* 10. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=203194 (visited on 03/02/2016).

— (2000). "Cooperation and Punishment in Public Goods Experiments". In: *American Economic Review* 90.4, pp. 980–994.

— (2002a). "Altruistic Punishment in Humans". In: *Nature* 415.10, pp. 137–140.

— (2002b). "Altruistic punishment in humans". In: *Nature* 415.6868, pp. 137–140. URL: https://www.nature.com/articles/415137a.

Fehr, Ernst et al. (2003). "A nation-wide laboratory: examining trust and trustworthiness by integrating behavioral experiments into representative survey". In: URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=385120 (visited on 01/13/2017).

Feinberg, Joel (1968). "Collective responsibility". In: *The Journal of Philosophy* 65.21, pp. 674–688.

Feldman, Daniel C. (1984). "The development and enforcement of group norms". In: *Academy of management review* 9.1, pp. 47–53. URL: http://amr.aom.org/content/9/1/47.short.

Fischbacher, Urs (2007a). "Z-Tree. Zurich Toolbox for Ready-made Economic Experiments". In: *Experimental Economics* 10.2. Published: University of Zurich, pp. 171–178.

— (2007b). "z-Tree: Zurich toolbox for ready-made economic experiments". en. In: *Experimental Economics* 10.2, pp. 171–178. ISSN: 1386-4157, 1573-6938. DOI: 10.1007/s10683-006-9159-4. URL: http://link.springer.com/article/10.1007/s10683-006-9159-4 (visited on 03/08/2016).

Forsyth, DR (2010). "Group dynamics . Wadsworth Cengage Learning". In: *Belmont, CA*.

French, Peter A. (1987). "Collective and corporate responsibility". In:

Fujimoto, Hiroaki and Eun-Soo Park (2010). "Framing effects and gender differences in voluntary public goods provision experiments". In: *The Journal of Socio-Economics* 39.4, pp. 455–457. ISSN: 1053-5357. DOI: 10.1016/j.socec.2010.03.002. URL: http://www.sciencedirect.com/science/article/pii/S1053535710000405 (visited on 01/05/2018).

Gaechter, Simon and Elke Renner (2010). "The effects of (incentivized) belief elicitation in public goods experiments". en. In: *Experimental Economics* 13.3, pp. 364–377. ISSN: 1386-4157, 1573-6938. DOI: 10.1007/s10683-010-9246-4. URL: http://link.springer.com/article/10.1007/s10683-010-9246-4 (visited on 03/23/2016).

Gaertner, Lowell and John Schopler (1998). "Perceived ingroup entitativity and intergroup bias: An interconnection of self and others". In: *European Journal of Social Psychology* 28.6, pp. 963–980.

Gellner, Ernest (2000). "Trust, cohesion, and the social order". In: *Trust: Making and breaking cooperative relations*, pp. 142–157. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.1742&rep=rep1&type=pdf (visited on 03/02/2016).

— (2008). *Nations and nationalism*. Cornell University Press.

Ginsburg, Tom and Alberto Simpser (2017). "Circles of Trust: A Proposal for Better Migrant Screening". In:

Glaeser, Edward L. et al. (2000). "Measuring Trust". In: *The Quarterly Journal of Economics* 115.3, pp. 811–846. ISSN: 0033-5533. URL: http://www.jstor.org/stable/2586897 (visited on 01/13/2017).

Gollwitzer, Mario and Livia Keller (2010). "What you did only matters if you are one of us". In: *Social Psychology*. URL: http://econtent.hogrefe.com/doi/full/10.1027/1864-9335/a000004 (visited on 10/09/2017).

Gould, Roger V. (1999). "Collective Violence and Group Solidarity: Evidence from a Feuding Society". In: *American Sociological Review* 64.3, pp. 356–380. ISSN: 0003-1224. DOI: 10.2307/2657491. URL: http://www.jstor.org/stable/2657491 (visited on 09/27/2017).

Grechenig, Kristoffel, Andreas Nicklisch, and Christian Thoeni (2010). "Punishment Despite Reasonable Doubt?A Public Goods Experiment with Sanctions Under Uncertainty". In: *Journal of Empirical Legal Studies* 7.4, pp. 847–867. ISSN: 1740-1461. DOI: 10.1111/j.1740-1461.2010.01197.x. URL: http://dx.doi.org/10.1111/j.1740-1461.2010.01197.x.

Greif, Avner (2002). "Institutions and impersonal exchange: from communal to individual responsibility". In: *Journal of Institutional and Theoretical Economics JITE* 158.1, pp. 168–204. URL: http://www.ingentaconnect.com/content/mohr/jite/2002/00000158/00000001/art00017 (visited on 09/27/2017).

Greif, Avner (2004). "Institutions and impersonal exchange: The European experience". In: URL: https://papers.ssrn.com/sol3/papers.cfm?abstract-id=548783 (visited on 09/27/2017).

Guala, Francesco (2009). "Methodological issues in experimental design and interpretation". In:

Guthrie, James P. (2000). "Alternative pay practices and employee turnover: An organization economics perspective". In: *Group & Organization Management* 25.4, pp. 419–439.

Gächter, Simon, Benedikt Herrmann, and Christian Thöni (2004). "Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence". In: *Journal of Economic Behavior & Organization*. Trust and Trustworthiness 55.4, pp. 505–531. ISSN: 0167-2681. DOI: 10.1016/j.jebo.2003.11.006. URL: http://www.sciencedirect.com/science/article/pii/S0167268104000708 (visited on 01/08/2018).

Haines III, Victor Y. and Simon Taggar (2006). "Antecedents of team reward attitude." In: *Group Dynamics: Theory, Research, and Practice* 10.3, p. 194.

Hardin (1968). "The tragedy of the commons". In: *Science* 162.3859, pp. 1243–1248.

Hardin, Russel (1995). *One for all*. Princeton: Princeton University Press.

Hawkes, Kristen and Rebecca Bliege (2002). "Showing off, handicap signaling, and the evolution of men's work". In: *Evolutionary Anthropology: Issues, News, and Reviews* 11.2, pp. 58–67.

Hechter, Michael (1988). *Principles of group solidarity*. Univ of California Press. URL: https://www.google.com/books?hl=ru&lr=&id=5rG670ZTGRQC&oi=fnd&pg=PP1&dq=Hechter,+M.,+1988.+Principles+of+Group+Solidarity.+University+of+California+Press.&ots=PMWO5RQXpa&sig=GQI8zp6hHgKoEU20-9xgUinUfKc (visited on 12/22/2016).

Heckathorn, Douglas D. (1988). "Collective sanctions and the creation of prisoner's dilemma norms". In: *American Journal of Sociology*, pp. 535–562. URL: http://www.jstor.org/stable/2780253 (visited on 03/02/2016).

— (1989). "Collective Action and the Second-Order Free-Rider Problem". In: *Rationality and Society* 1, pp. 78–100.

— (1990). "Collective Sanctions and Compliance Norms: A Formal Theory of Group-Mediated Social Control". In: *American Sociological Review* 55.3, pp. 366–384. ISSN: 00031224. URL: http://www.jstor.org/stable/2095762.

Hennig-Schmidt, Heike and Ulrike Leopold-Wildburger (2014). "The shadow of the past: how experience affects behavior in an iterated prisoner's dilemma experiment". en. In: *Journal of Business Economics* 84.6, pp. 865–878. ISSN: 0044-2372, 1861-8928. DOI: 10.1007/s11573-014-0707-7. URL: https://link.springer.com/article/10.1007/s11573-014-0707-7 (visited on 09/24/2017).

Herrmann, Benedikt, Christian Thöni, and Simon Gächter (2008). "Antisocial punishment across societies". In: *Science* 319.5868, pp. 1362–1367. URL: http://science.sciencemag.org/content/319/5868/1362.short (visited on 12/09/2016).

Hisamova, Zarina (2002). "Chto Podumaet Sosed Vassily [What my neightbor Vassily would think about it?]" In: *Expert* 38.

Homans, George Caspar (2017). *The human group*. Routledge.

Hopfensitz, Astrid and Ernesto Reuben (2009). "The importance of emotions for the effectiveness of social punishment". In: *The Economic Journal* 119.540, pp. 1534–1559. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0297.2009.02288.x/full.

Horowitz, Donald L. (1985). *Ethnic Groups in Conflict*. en. Google-Books-ID: Q82saX1HVQYC. University of California Press. ISBN: 978-0-520-05385-4.

Horvitz, Leslie Alan and Christopher Catherwood (2009). *Encyclopedia of war crimes and genocide*. Infobase Publishing. URL: https://books.google.com/books?hl=en&lr=&id=AHpFp2nsGyUC&oi=fnd&pg=PR3&dq=20.%09Horvitz,+Leslie+Alan,+and+Christopher+Catherwood.+2009.+Encyclopedia+of+war+crimes+and+genocide.+Infobase+Publishing.&ots=MxlRwUHcO-&sig=_HNfFy763nrVxpefx4Ho3t2eXC0 (visited on 03/02/2016).

Ibrahim, Darian M. (2008). "Individual or Collective Liability for Corporate Directors?" In: *Iowa Law Review* 93, p. 929. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=918119 (visited on 03/02/2016).

Jaspers, Karl (2009). *The question of German guilt*. 16. Fordham Univ Press.

Joseph, Joshua (2003). *Ethics Resource Center National Business Ethics Survey (2003): How Employees View Ethics in Their Organizations*. Ethics Resource Center.

Kahneman, Daniel and Amos Tversky (2013). "Prospect theory: An analysis of decision under risk". In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, pp. 99–127.

Kalyvas, Stathis N (2004). "The paradox of terrorism in civil war". In: *The Journal of Ethics* 8.1, pp. 97–138.

— (2006). *The logic of violence in civil war*. Cambridge University Press.

Kandori, Michihiro (1992). "Social norms and community enforcement". In: *The Review of Economic Studies* 59.1, pp. 63–80.

Kerr, John, Mamta Vardhan, and Rohit Jindal (2012). "Prosocial behavior and incentives: evidence from field experiments in rural Mexico and Tanzania". In: *Ecological Economics* 73, pp. 220–227.

Kerrin, Máire and Nick Oliver (2002). "Collective and individual improvement activities: the role of reward systems". In: *Personnel review* 31.3, pp. 320–337. URL: http://www.emeraldinsight.com/doi/abs/10.1108/00483480210422732 (visited on 03/02/2016).

Kimbrough, Erik O. and Jared Rubin (2015). "Sustaining group reputation". In: *The Journal of Law, Economics, and Organization* 31.3, pp. 599–628. URL: https://academic.oup.com/jleo/article-abstract/31/3/599/808400 (visited on 09/30/2017).

Kokko, Hanna, Rufus A. Johnstone, and Tim H. Clutton-Brock (2001). "The evolution of cooperative breeding through group augmentation". In: *Proceedings of the Royal Society of London B: Biological Sciences* 268.1463, pp. 187–196.

Kollock, Peter (1998). "Social dilemmas: The anatomy of cooperation". In: *Annual review of sociology* 24.1, pp. 183–214.

Kraakman, Reinier H. (1986). "Gatekeepers: The Anatomy of a Third-Party Enforcement Strategy". In: *Journal of Law, Economics, & Organization* 2.1, pp. 53–104. ISSN: 8756-6222. URL: http://www.jstor.org/stable/764916 (visited on 09/24/2017).

Kruglanski, Arie W., Irith Friedman, and Gabriella Zeevi (1971). "The effects of extrinsic incentive on some qualitative aspects of task performance". In: *Journal of Personality* 39.4, pp. 606–617. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-6494.1971.tb00066.x/full.

Landa, Janet T. (1994). *Trust, Ethnicity, and Identity: Beyond the New Institutional Economics of Ethnic Trading Networks, Contract Law, and Gift-exchange*. en. Google-Books-ID: 0ZHxmvDJq40C. University of Michigan Press. ISBN: 978-0-472-10361-4.

Le Bon, Gustave (1897). *The crowd: A study of the popular mind*. Fischer.

Ledyard, John O. (1994). "Public goods: A survey of experimental research". In:

Leeson, Peter T. (2008). "Social distance and self-enforcing exchange". In: *The Journal of Legal Studies* 37.1, pp. 161–188. URL: http://www.journals.uchicago.edu/doi/abs/10.1086/588262 (visited on 10/02/2017).

Lergetporer, Philipp et al. (2014). "Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation". In: *Proceedings of the National Academy of Sciences* 111.19, pp. 6916–6921.

Levinson, Daryl J. (2003). "Collective sanctions". In: *Stanford Law Review*, pp. 345–428. URL: http://www.jstor.org/stable/1229612 (visited on 03/02/2016).

Lickel, Brian, Toni Schmader, and David L. Hamilton (2003). "A case of collective responsibility: Who else was to blame for the Columbine High School shootings?" In: *Personality and Social Psychology Bulletin* 29.2, pp. 194–204. URL: http://journals.sagepub.com/doi/abs/10.1177/0146167202239045 (visited on 09/30/2017).

Lickel, Brian et al. (2006). "Vicarious retribution: The role of collective blame in intergroup aggression". In: *Personality and Social Psychology Review* 10.4, pp. 372–390. URL: http://journals.sagepub.com/doi/abs/10.1207/s15327957pspr1004_6 (visited on 09/30/2017).

Lieberman, Debra and Lance Linke (2007). "The effect of social category on third party punishment". In: *Evolutionary Psychology* 5.2, p. 147470490700500203.

Lind, E. Allan and Tom R. Tyler (1988). *The social psychology of procedural justice*. Springer Science & Business Media. URL: https://books.google.ch/books?hl=en&lr=&id=oyXZ5IM0J8MC&oi=fnd&pg=PA1&dq=.%C2%A0The+social+psychology+of+procedural+justice&ots=QsECzTE16X&sig=YOdNbstjymP4tErTmawB8rgW3-M.

Locke, John (2007). *Some Thoughts Concerning Education:(Including Of the Conduct of the Understanding)*. Courier Corporation. URL: https://books.google.ch/books?hl=en&lr=&id=l0c1AwAAQBAJ&oi=fnd&pg=PP1&dq=Some+Thoughts+Concerning+Education&ots=5qWEFZ5RVS&sig=Hep_89Hom8nQ6aJwWVUj4Gk8aAg.

Loewenstein, George F. et al. (2001). "Risk as feelings." In: *Psychological bulletin* 127.2, p. 267. URL: http://sci-hub.cc/http://psycnet.apa.org/journals/bul/127/2/267/.

Lucas, Jeffrey W. (2003). "Theory-Testing, Generalization, and the Problem of External Validity". In: *Sociological Theory* 21.3, pp. 236–253. ISSN: 07352751. URL: http://www.jstor.org/stable/3108637.

Maas, Victor S. and Marcel Van Rinsum (2013). "How Control System Design Influences Performance Misreporting". en. In: *Journal of Accounting Research* 51.5, pp. 1159–1186. ISSN: 1475-679X. DOI: 10.1111/1475-679X.12025. URL: http://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12025/abstract (visited on 01/23/2018).

Mannheim, Hermann (2013). *Group problems in crime and punishment*. Routledge. URL: https://books.google.com/books?hl=en&lr=&id=fS-BAAAAQBAJ&oi=fnd&pg=PP1&dq=26.%09Mannheim,+Hermann.+2013.+Group+problems+in+crime+and+punishment.+Routledge&ots=eEObIGgj9C&sig=6_Qfg7V3EvvCxYwSZwCj4vRM4VY (visited on 03/02/2016).

Marcus-Newhall, Amy et al. (2000). "Displaced aggression is alive and well: a meta-analytic review." In: *Journal of personality and social psychology* 78.4, p. 670.

Marques, José M., Vincent Y. Yzerbyt, and Jacques-Philippe Leyens (1988). "The "Black Sheep Effect": Extremity of judgments towards ingroup members as a function of group identification". en. In: *European Journal of Social Psychology* 18.1, pp. 1–16. ISSN: 1099-0992. DOI: 10.1002/ejsp.2420180102. URL: http://onlinelibrary.wiley.com/doi/10.1002/ejsp.2420180102/full (visited on 01/08/2018).

Masclet, David and Marie-Claire Villeval (2008). "Punishment, inequality, and welfare: a public good experiment". In: *Social Choice and Welfare* 31.3, pp. 475–502.

Matthews, David (2016). *'We are tough': a rector's fight against corruption in Kazakhstan*. URL: https://www.timeshighereducation.com/news/we-are-tough-rectors-fight-against-corruption-kazakhstan (visited on 12/24/2016).

May, Larry (2006). "State aggression, collective liability, and individual mens rea". In: *Midwest studies in philosophy* 30.1, pp. 309–324.

McAuliffe, Katherine, Jillian J. Jordan, and Felix Warneken (2015). "Costly third-party punishment in young children". In: *Cognition* 134, pp. 1–10.

Mockenhaupt, Brian (2007). "The Army We Have". In: *The Atlantic*. ISSN: 1072-7825. URL: http://www.theatlantic.com/magazine/archive/2007/06/the-army-we-have/305902/ (visited on 12/27/2016).

Monahan, John and Laurens Walker (2011). "Twenty-Five Years of <i>Social Science in Law</i>" in: *Law and Human Behavior* 35.1, pp. 72–82. ISSN: 0147-7307. URL: http://dx.doi.org/10.1007/s10979-009-9214-8.

Mulder, Laetitia B. and Rob M. A. Nelissen (2010). "When Rules Really Make a Difference: The Effect of Cooperation Rules and Self-Sacrificing Leadership on Moral Norms in Social Dilemmas". en. In: *Journal of Business Ethics* 95.1, pp. 57–72. ISSN: 0167-4544, 1573-0697. DOI: 10.1007/s10551-011-0795-z. URL: https://link.springer.com/article/10.1007/s10551-011-0795-z (visited on 01/27/2018).

Mulder, Laetitia B. et al. (2005). "The effect of feedback on support for a sanctioning system in a social dilemma: The difference between installing and maintaining the sanction". In: *Journal of Economic Psychology* 26.3, pp. 443–458. URL: http://www.sciencedirect.com/science/article/pii/S0167487004001199.

Nakao, Keisuke and Sun-Ki Chai (2011). "Criminal conflict as collective punishment". In: *Economics of Peace and Security Journal* 6.1, pp. 5–11. URL: https://ideas.repec.org/a/epc/journl/v6y2011i1p5-11.html.

Narloch, Ulf, Unai Pascual, and Adam G Drucker (2012). "Collective action dynamics under external rewards: experimental insights from Andean farming communities". In: *World Development* 40.10, pp. 2096–2107.

Narveson, Jan (2002). "Collective responsibility". In: *The Journal of Ethics* 6.2, pp. 179–198.

Nelissen, Rob MA and Marcel Zeelenberg (2009). "Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions". In: *Judgment and Decision making* 4.7, p. 543.

Neuman, W. Lawrence (2013). *Social research methods: Qualitative and quantitative approaches*. Pearson education.

Newheiser, Anna-Kaisa and John F. Dovidio (2015). "High outgroup entitativity can inhibit intergroup retribution". In: *British Journal of Social Psychology* 54.2, pp. 341–358.

Newheiser, Anna-Kaisa, Takuya Sawaoka, and John F. Dovidio (2012). "Why do we punish groups? High entitativity promotes moral suspicion". In: *Journal of Experimental Social Psychology* 48.4, pp. 931–936. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2012.02.013. URL: http://www.sciencedirect.com/science/article/pii/S002210311200025X (visited on 10/24/2017).

Nikiforakis, Nikos (2008). "Punishment and counter-punishment in public good games: Can we really govern ourselves?" In: *Journal of Public Economics* 92.1, pp. 91–112. ISSN: 0047-2727. DOI: 10.1016/j.jpubeco.2007.04.008. URL: http://www.sciencedirect.com/science/article/pii/S0047272707000643 (visited on 10/24/2017).

Nikiforakis, Nikos and Hans-Theo Normann (2008). "A comparative statics analysis of punishment in public-good experiments". In: *Experimental Economics* 11.4, pp. 358–369. URL: http://www.springerlink.com/index/971H6R48182358KW.pdf (visited on 03/02/2016).

Normann, Hans-Theo and Brian Wallace (2012). "The impact of the termination rule on cooperation in a prisoner's dilemma experiment". en. In: *International Journal of Game Theory* 41.3, pp. 707–718. ISSN: 0020-7276, 1432-1270. DOI: 10.1007/s00182-012-0341-y. URL: https://link.springer.com/article/10.1007/s00182-012-0341-y (visited on 09/30/2017).

Nowak, Martin A. and Karl Sigmund (2005). "Evolution of indirect reciprocity". In: *Nature* 437.7063, pp. 1291–1298. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature04131.

Ostrom, Elinor (2015). *Governing the commons*. Cambridge university press.

Page, Talbot, Louis Putterman, and Bulent Unel (2005). "Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency". In: *The Economic Journal* 115.506, pp. 1032–1053.

Pereira, Andrea et al. (2015). "Collective punishment depends on collective responsibility and political organization of the target group". In: *Journal of Experimental Social Psychology* 56, pp. 4–17. URL: http://www.sciencedirect.com/science/article/pii/S0022103114001267 (visited on 03/02/2016).

Peterson, Laura (2012). "Collective Sanctions: Learning from the NFL's Justifiable Use of Group Punishment". In: *Texas Review of Entertainment & Sports Law* 14, p. 165. URL: http://heinonline.org/HOL/Page?handle=hein.journals/tresl14&id=185&div=&collection=.

Pettigrew, Thomas F. and Linda R. Tropp (2006). "A meta-analytic test of intergroup contact theory." In: *Journal of personality and social psychology* 90.5, p. 751.

Pettit, Philip (2007). "Responsibility incorporated". In: *Ethics* 117.2, pp. 171–201. URL: http://www.journals.uchicago.edu/doi/full/10.1086/510695.

Piaget, Jean (1965). "The moral judgment of the child (1932)". In: *New York: The Free*.

Prooijen, Jan-Willem, Marcello Gallucci, and Gaby Toeset (2008). "Procedural justice in punishment systems: Inconsistent punishment procedures have detrimental effects on cooperation". In: *British Journal of Social Psychology* 47.2, pp. 311–324.

Rand, David G. and Martin A. Nowak (2011). "The evolution of anti-social punishment in optional public goods games". In: *Nature communications* 2, p. 434.

Rosenthal, Abraham Michael (2015). *Thirty-eight witnesses: The Kitty Genovese case*. Open Road Media.

Ross, H. Laurence (1975). "The Scandinavian Myth: The Effectiveness of Drinking and Driving Legislation in Sweden and Norway". In: *Journal ofLegal Studies* 4, pp. 285–310.

Rönnegard, David (2015). *The Fallacy of Corporate Moral Agency*. Springer. URL: http://link.springer.com/content/pdf/10.1007/978-94-017-9756-6.pdf.

Sally, David (1995). "Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992". In: *Rationality and society* 7.1, pp. 58–92.

Sasaki, Tatsuya and Satoshi Uchida (2013). "The evolution of cooperation by social exclusion". In: *Proc. R. Soc. B*. Vol. 280. The Royal Society, p. 20122498.

Schiappa, Edward, Peter B Gregg, and Dean E Hewes (2005). "The parasocial contact hypothesis". In: *Communication monographs* 72.1, pp. 92–115.

Schroeder, David A. (1995). *Social dilemmas: Perspectives on individuals and groups*. Greenwood Publishing Group. URL: https://books.google.ch/books?hl=en&lr=&id=4lauxt-yLTgC&oi=fnd&pg=PP5&dq=+Schroeder,+D.+A.+ed.,+1995.+%C2%A0Social+dilemmas:+Perspectives+on+individuals+and+groups.+&ots=F2jx0TW6_a&sig=vlrJtziyEeaOFXOEu2IVEIfx2AA.

Schuck, Glenn (2010). *Bathroom Ban Leads To Riot At NYC High School*. URL: http://newyork.cbslocal.com/2010/12/10/bathroom-ban-leads-to-riot-at-nyc-high-school/ (visited on 12/27/2016).

Seguino, Stephanie, Thomas Stevens, and Mark Lutz (1996). "Gender and cooperative behavior: economic man rides alone". In: *Feminist Economics* 2.1, pp. 1–21. URL: http://www.tandfonline.com/doi/abs/10.1080/738552683 (visited on 03/02/2016).

Sell, Jane, W. I. Griffith, and Rick K. Wilson (1993). "Are Women More Cooperative Than Men in Social Dilemmas?" In: *Social Psychology Quarterly* 56.3, pp. 211–222. ISSN: 01902725. URL: http://www.jstor.org/stable/2786779.

Sherif, Muzafer (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment*. Vol. 10. University Book Exchange Norman, OK. URL: http://psychclassics.yorku.ca/Sherif/chap7.htm?wptouch_preview_theme=enabled (visited on 09/27/2017).

Shinada, M., T. Yamagishi, and Y. Ohmura (2004). "False friends are worse than bitter enemies: "Altruistic" punishment of in-group members". In: *Evolution and Human Behavior* 25.6, pp. 379–393. URL: <GotoISI>://000225275300003.

Silverman, Irwin and Marion Eals (1992). "Sex differences in spatial abilities: evolutionary theory and data." In: *Portions of this paper were presented at the meetings of the International Society for Human Ethology in Binghamton, NY, Jun 1990, the Human Behavior and Evolution Society in Los Angeles, CA, Aug 1990, and the European Sociobiological Society in Prague, Czechoslovakia, Aug 1991*. Oxford University Press.

Simmel, George (2010). *Conflict and the web of group affiliations*. Simon and Schuster.

Singer, Tania and Nikolaus Steinbeis (2009). "Differential Roles of Fairness- and Compassion-Based Motivations for Cooperation, Defection, and Punishment". en. In: *Annals of the New York Academy of Sciences* 1167.1, pp. 41–

50. ISSN: 1749-6632. DOI: 10.1111/j.1749-6632.2009.04733.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.2009.04733.x/abstract (visited on 09/24/2017).

Sjöström, Arne and Mario Gollwitzer (2015). "Displaced revenge: Can revenge taste "sweet" if it aims at a different target?" In: *Journal of Experimental Social Psychology* 56, pp. 191–202. URL: http://www.sciencedirect.com/science/article/pii/S0022103114001528 (visited on 09/30/2017).

Small, Deborah A. and George Loewenstein (2005). "The devil you know: The effects of identifiability on punishment". In: *Journal of Behavioral Decision Making* 18.5, pp. 311–318. URL: http://onlinelibrary.wiley.com/doi/10.1002/bdm.507/full (visited on 10/09/2017).

Solow, John L. and Nicole Kirkwood (2002). "Group identity and gender in public goods experiments". In: *Journal of Economic Behavior & Organization* 48.4, pp. 403–412. URL: http://www.sciencedirect.com/science/article/pii/S0167268101002438 (visited on 03/02/2016).

Stenstrom, Douglas M. et al. (2008). "The roles of ingroup identification and outgroup entitativity in intergroup retribution". In: *Personality and Social Psychology Bulletin* 34.11, pp. 1570–1582.

Stoff, Christian (2006). *Establishing cooperation between groups: Ingroup versus outgroup punishment*. Tech. rep. Working Paper, Socioeconomic Institute, University of Zurich. URL: https://www.econstor.eu/handle/10419/76148.

Stürmer, Stefan and Mark Snyder (2009). *The Psychology of Prosocial Behavior: Group Processes, Intergroup Relations, and Helping*. en. Google-Books-ID: gGOlpbBUgiMC. John Wiley & Sons. ISBN: 978-1-4443-0795-5.

Sumner, William Graham (2013). *Folkways-A Study Of The Sociological Importance Of Usages, Manners, Customs, Mores And Morals*. Read Books Ltd.

Tajfel, Henri (1970). "Experiments in intergroup discrimination". In: *Scientific American* 223.5, pp. 96–103. URL: http://www.jstor.org/stable/24927662.

— (1982). "Social Psychology of intergroup relations". In: *Annual Review of Psychology* 33, pp. 1–39.

Tajfel, Henri and J.C. Turner (1979). "An integrative theory of social contact." In: *The Social psychology of intergroup relations*. Ed. by William G. Austin and Stephen Worchel. Monterey, Calif.: Brooks/Cole Pub. Co., pp. 162–173. ISBN: 0-8185-0278-9.

Tenbrunsel, Ann E., Kristin Smith-Crowe, and Elizabeth E. Umphress (2003). "Building houses on rocks: The role of the ethical infrastructure in organizations". In: *Social justice research* 16.3, pp. 285–307. URL: http://www.springerlink.com/index/Q0X2M35385574547.pdf.

Thöni, Christian, Jean-Robert Tyran, and Erik Wengström (2012). "Microfoundations of social capital". In: *Journal of Public Economics* 96.7–8, pp. 635–643. ISSN: 0047-2727. DOI: 10.1016/j.jpubeco.2012.04.003. URL: http://www.sciencedirect.com/science/article/pii/S0047272712000370 (visited on 01/04/2017).

Travis, Jeremy, Bruce Western, and F. Stevens Redburn (2014). "The growth of incarceration in the United States: Exploring causes and consequences". In: URL: http://academicworks.cuny.edu/jj_pubs/27/.

Trevino, Linda Klebe and Bart Victor (1992). "Peer reporting of unethical behavior: A social context perspective". In: *Academy of Management Journal* 35.1, pp. 38–64. URL: http://amj.aom.org/content/35/1/38.short.

Turner, John C. and Penelope J. Oakes (1986). "The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence". In: *British Journal of Social Psychology* 25.3, pp. 237–252.

Varian, Hal R. (1990). "Monitoring Agents With Other Agents". In: *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft* 146.1, pp. 153–174. ISSN: 0932-4569. URL: http://www.jstor.org/stable/40751313 (visited on 01/06/2018).

Velikonja, Urska (2011). "Leverage, sanctions, and deterrence of accounting fraud". In: *UC Davis Law Review* 44, pp. 1281–1345. URL: http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1651314 (visited on 03/02/2016).

Webster, Murray and Jane Sell (2014). *Laboratory experiments in the social sciences*. Elsevier.

Whitmeyer, Joseph M. (2002). "The compliance you need for a cost you can afford: How to use individual and collective sanctions?" In: *Social Science Research* 31.4, pp. 630–652. URL: http://www.sciencedirect.com/science/article/pii/S0049089X02000170 (visited on 03/02/2016).

Whittaker, Hannah (2015). "Legacies of Empire: State Violence and Collective Punishment in Kenya's North Eastern Province, c. 1963–Present". In: *The Journal of Imperial and Commonwealth History* 43.4, pp. 641–657.

Wilson, Monica Hunter (1987). *Good company: A study of Nyakyusa age-villages*. Waveland PressInc.

Wilson, Steven Lloyd (2015). "Social identity, cross-cutting cleavages, and explaining the breakdown of interethnic cooperation". en. In: *Rationality and Society* 27.4, pp. 455–468. ISSN: 1043-4631. DOI: 10.1177/1043463115605301. URL: https://doi.org/10.1177/1043463115605301 (visited on 01/09/2018).

Wright, Bradley RE et al. (2004). "Does the perceived risk of punishment deter criminally prone individuals? Rational choice, self-control, and crime". In: *Journal of Research in Crime and Delinquency* 41.2, pp. 180–213.

Yamagishi, Toshio and Toko Kiyonari (2000). "The group as the container of generalized reciprocity". In: *Social Psychology Quarterly*, pp. 116–132. URL: http://www.jstor.org/stable/2695887 (visited on 12/09/2016).

Yefeng, Chen, Shu-Guang Jiang, and Marie Claire Villeval (2015). "The Tragedy of Corruption. Corruption as a Social Dilemma". In: *Corruption as a Social Dilemma (December 1, 2015)*. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2698050 (visited on 12/22/2016).

Zelmer, Jennifer (2003). "Linear public goods experiments: A meta-analysis". In: *Experimental Economics* 6.3, pp. 299–310. URL: http://link.springer.com/article/10.1023/A:1026277420119 (visited on 03/02/2016).